$$P(COVID = \mathrm{T} \mid Mask = \mathrm{F}, Social\ Distancing = \mathrm{T}) = ?$$

Data sampled from 12 individuals

#	COVID	Mask	Social Distancing
1	True	False	True
2	False	True	True
3	True	False	False
4	False	True	True
5	False	True	True
6	False	True	False
7	True	False	True
8	False	True	True
9	False	True	True
10	False	False	False
11	False	True	False
12	True	True	True
Total	4 (False)		
	3 (True,False)		

How to answer an inquiry from data? Let's start simple:

$$P(\underline{C}OVID = \mathtt{T} \mid \underline{M}ask = \mathtt{F}) = ?$$

- Using **conditional probability**,

$$P(\underline{C}OVID = \mathtt{T} \mid \underline{M}ask = \mathtt{F}) = \frac{P(COVID{=}\mathtt{T}, Mask{=}\mathtt{F})}{P(Mask{=}\mathtt{F})}$$

- $P(COVID = \mathtt{T}, Mask = \mathtt{F}) = \dfrac{\#(True, False)}{12} = \dfrac{3}{12}$
- Using **marginal probability**,

$$P(Mask = \mathtt{F}) = \frac{\#(*, False)}{12} = \frac{4}{12}$$

- Hence,

$$P(\underline{C}OVID = \mathtt{T} \mid \underline{M}ask = \mathtt{F}) = \frac{3/12}{4/12} = 0.75$$

Data sampled from 12 individuals

#	COVID	Mask	Social Distancing
1	True	False	True
2	False	True	True
3	True	False	False
4	False	True	True
5	False	True	True
6	False	True	False
7	True	False	True
8	False	True	True
9	False	True	True
10	False	False	False
11	False	True	False
12	True	True	True
Total	4 (False)		
	3 (True,False)		

- What about

$$P(\underline{C}OVID = \mathtt{F} \mid \underline{M}ask = \mathtt{T}, Social\ \underline{D}istancing = \mathtt{T}) = ?$$

- Using **conditional probability**,

$$P(C = \mathtt{F} \mid M = \mathtt{T}, D = \mathtt{T}) = \frac{P(C = \mathtt{F}, M = \mathtt{T}, D = \mathtt{T})}{P(M = \mathtt{T}, D = \mathtt{T})} = \frac{5/12}{}$$

#	COVID	Mask	Social Distancing
1	True	False	True
2	False	True	True
3	True	False	False
4	False	True	True
5	False	True	True
6	False	True	False
7	True	False	True
8	False	True	True
9	False	True	True
10	False	False	False
11	False	True	False
12	True	True	True

- What about

$$P(\underline{C}OVID = \mathtt{F} \mid \underline{M}ask = \mathtt{T}, Social\ \underline{D}istancing = \mathtt{T}) = ?$$

- Using **conditional probability**,

$$P(C = \mathtt{F} \mid M = \mathtt{T}, D = \mathtt{T}) = \frac{P(C = \mathtt{F}, M = \mathtt{T}, D = \mathtt{T})}{P(M = \mathtt{T}, D = \mathtt{T})} = \frac{5/12}{6/12}$$

- Using **marginal probability**,

$$P(M = \mathtt{T}, D = \mathtt{T}) = \sum_{C=\mathtt{T},\mathtt{F}} P(C, M = \mathtt{T}, D = \mathtt{T})$$

$$= 6/12$$

#	COVID	Mask	Social Distancing
1	True	False	True
2	False	True	True
3	True	False	False
4	False	True	True
5	False	True	True
6	False	True	False
7	True	False	True
8	False	True	True
9	False	True	True
10	False	False	False
11	False	True	False
12	True	True	True

- Now, what about

$$P(C = \mathtt{F}, M \neq \mathtt{T} \mid D = \mathtt{T}) = ?$$

$$P(M = \mathtt{T} \mid C = \mathtt{F}, D = \mathtt{F}) = ?$$

- The key is to find the *joint probability distribution* for C, M, and D, i.e.,

$$P(C, M, D)$$

with 8 parameters $p_1, \cdots, p_8$, where $\sum_{i=1}^{8} p_i = 1$.

- So the number of parameters to obtain $P(C, M, D)$ is $2^3 - 1 = 7$.

Joint distribution

C	M	D	$P(C, M, D)$
False	False	False	$p_1 = 1/12$
False	False	True	$p_2 = 0$
False	True	False	$p_3 = 2/12$
False	True	True	$p_4 = 5/12$
True	False	False	$p_5 = 1/12$
True	False	True	$p_6 = 2/12$
True	True	False	$p_7 = 0$
True	True	True	$p_8 = 1/12$

- The key is to find the *joint probability distribution* for C, M, and D, i.e.,

$$P(C, M, D)$$

with 8 parameters $p_1, \cdots, p_8$, where $\sum_{i=1}^{8} p_i = 1$.

- So the number of parameters to obtain $P(C, M, D)$ is $2^3 - 1 = 7$.

Joint distribution

C	M	D	$P(C, M, D)$
False	False	False	$p_1 = 1/12$
False	False	True	$p_2 = 0$
False	True	False	$p_3 = 2/12$
False	True	True	$p_4 = 5/12$
True	False	False	$p_5 = 1/12$
True	False	True	$p_6 = 2/12$
True	True	False	$p_7 = 0$
True	True	True	$p_8 = 1/12$

The role of joint distribution

If we have the **joint distribution**, we can calculate any probability using the **conditional distribution** and **marginal distribution**.

So, we can look for the **joint distribution** to answer inquiries.

Given random variables $X_1, X_2, \ldots, X_n$,

- **joint probability distribution**:

$$P(X_1, X_2, \ldots, X_n)$$

- **marginal probability distribution**:

$$P(X_1) = \sum_{X_2, \ldots, X_n} P(X_1, X_2, \ldots, X_n)$$

- **conditional probability distribution (CPD)**:

$$P(X_1 \mid X_2, \ldots, X_n) = \frac{P(X_1, X_2, \ldots, X_n)}{P(X_2, \ldots, X_n)}$$

where the denominator is a marginal distribution:

$$P(X_2, \ldots, X_n) = \sum_{X_1} P(X_1, X_2, \ldots, X_n)$$

- The goal is to find from data, the joint probability distribution.

$P(C, M, D) = ?$

#	C	M	D
1	True	False	True
2	False	True	True
3	True	False	False
4	False	True	True
5	False	True	True
6	False	True	False
7	True	False	True
8	False	True	True
9	False	True	True
10	False	False	False
11	False	True	False
12	True	True	True

#	COVID	Mask	Social Distancing	FIU	COugh	Fever	Ventilation	Season	ConGestion	Difficulty Breathing	DRug	Allergy
1	True	False	True	False	True	True	True	Spring	True	True	False	False
2	False	True	True	False	False	True	False	Summer	False	False	True	False
3	True	False	False	True	True	False	False	Fall	False	True	True	False
4	False	True	True	False	False	True	False	Winter	True	True	False	True
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1000	True	True	True	False	True	False	True	Spring	False	False	True	True

$$P(COVID, Mask, Social\ Distancing, ..., DRug, Allergy)$$

- 12 variables, all binary except for Season which takes 4 values.

#	COVID	Mask	Social Distancing	FIU	COugh	Fever	Ventilation	Season	ConGestion	Difficulty Breathing	DRug	Allergy
1	True	False	True	False	True	True	True	Spring	True	True	False	False
2	False	True	True	False	False	True	False	Summer	False	False	True	False
3	True	False	False	True	True	False	False	Fall	False	True	True	False
4	False	True	True	False	False	True	False	Winter	True	True	False	True
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1000	True	True	True	False	True	False	True	Spring	False	False	True	True

$$P(COVID, Mask, Social\ Distancing, ..., DRug, Allergy)$$

- 12 variables, all binary except for Season which takes 4 values.
- Number of required parameters is $2^{11} \times 4 - 1 = 8191$,
- whereas we only have 1000 data instances, so at least 7191 parameters will be set to zero? The fact that an instance does not appear in our data, does not mean that it never (or even rarely) happens.

Statistical Independence

- For **C**OVID, **M**ask, Social **D**istancing, using conditional probability,

$$P(C, M, D) = P(C \mid M, D)P(M, D).$$

- Suppose M and D are **(statistically/probabilistically/mutually) independent**, denoted $M \perp D$, i.e.,

$$P(M, D) = P(M)P(D).$$

- Then

$$P(C, M, D) = P(C \mid M, D)P(M)P(D).$$

Number of parameters $= 2^2 + (2 - 1) + (2 - 1) = 6$.

The importance of independence

Independencies can reduce the number of parameters.

- Random variables X and Y are **independent**, i.e., $X \perp Y$, if

$$P(X, Y) = P(X)P(Y)$$

or equivalently

$$P(X \mid Y) = P(X) \quad \text{or} \quad P(Y \mid X) = P(Y).$$

- So they are **not independent** if and only if there exists two values of Y, say y_1 and y_2, such that

$$P(X \mid Y = y_1) \neq P(X \mid Y = y_2).$$

Back to the COVID problem with 12 variables:

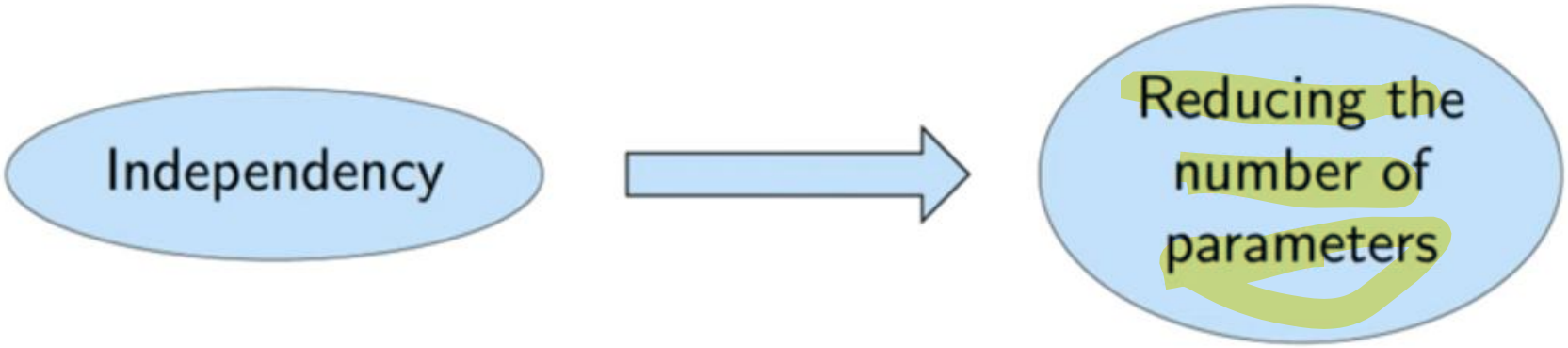
#	COVID	Mask	Social Distancing	FIU	COugh	Fever	Ventilation	Season	ConGestion	Difficulty Breathing	DRug	Allergy
1	True	False	True	False	True	True	True	Spring	True	True	False	False
2	False	True	True	False	False	True	False	Summer	False	False	True	False
3	True	False	False	True	True	False	False	Fall	False	True	True	False
4	False	True	True	False	False	True	False	Winter	True	True	False	True
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1000	True	True	True	False	True	False	True	Spring	False	False	True	True

- Suppose all 12 random variables are mutually independent:

$$C \perp M,\ C \perp D,\ \ldots, R \perp A$$

$$P(C, M, D, U, ..., A) = P(C)P(M)P(D)...P(A)$$

- Number of parameters for the joint distribution $= 11+3 = 14$.
- The previous number was 8191!



Independence is, however, not ubiquitous. For example, the independence assumption in the COVID problem may be wrong:

$$Mask \not\perp Social\ distancing$$

Because some people who wear a mask may no longer keep a social distance as they may think they are safe.

Conditional independence: The COVID example

Consider **F**ever and **P**CR Test Result.

- If someone has fever, we can guess their test result and vice versa:

$$P(F = 1 \mid P = 0) = 0.2 \quad \text{and} \quad P(F = 1 \mid P = 1) = 0.7$$

Conditional independence: The COVID example

Consider **F**ever and **P**CR Test Result.

- If someone has fever, we can guess their test result and vice versa:

$$P(F = 1 \mid P = 0) = 0.2 \quad \text{and} \quad P(F = 1 \mid P = 1) = 0.7$$

$$P(P = 1 \mid F = 0) = 0.3 \quad \text{and} \quad P(P = 1 \mid F = 1) = 0.4$$

Conditional independence: The COVID example

Consider **F**ever and **P**CR Test Result.

- If someone has fever, we can guess their test result and vice versa:

$$P(F = 1 \mid P = 0) = 0.2 \quad \text{and} \quad P(F = 1 \mid P = 1) = 0.7$$
$$P(P = 1 \mid F = 0) = 0.3 \quad \text{and} \quad P(P = 1 \mid F = 1) = 0.4$$

$$\Longrightarrow \text{Fever and PCR are } \textbf{not} \text{ independent.}$$
$$F \not\perp P$$

- If we know the person has **C**OVID ($C = 1$), they have fever and test positive with a high probability:

Conditional independence: The COVID example

Consider **F**ever and **P**CR Test Result.

- If someone has fever, we can guess their test result and vice versa:

$$P(F = 1 \mid P = 0) = 0.2 \quad \text{and} \quad P(F = 1 \mid P = 1) = 0.7$$
$$P(P = 1 \mid F = 0) = 0.3 \quad \text{and} \quad P(P = 1 \mid F = 1) = 0.4$$

$$\Longrightarrow \text{Fever and PCR are } \textbf{not} \text{ independent.}$$
$$F \not\perp P$$

- If we know the person has **C**OVID ($C = 1$), they have fever and test positive with a high probability:

$$P(F = 1 \mid C = 1) = 0.6 \quad \text{and} \quad P(P = 1 \mid C = 1) = 0.9$$

Conditional independence: The COVID example

Consider **F**ever and **P**CR Test Result.

- If someone has fever, we can guess their test result and vice versa:

$$P(F = 1 \mid P = 0) = 0.2 \quad \text{and} \quad P(F = 1 \mid P = 1) = 0.7$$
$$P(P = 1 \mid F = 0) = 0.3 \quad \text{and} \quad P(P = 1 \mid F = 1) = 0.4$$

$$\Longrightarrow \text{Fever and PCR are } \textbf{not} \text{ independent.}$$
$$F \not\perp P$$

- If we know the person has **C**OVID ($C = 1$), they have fever and test positive with a high probability:

$$P(F = 1 \mid C = 1) = 0.6 \quad \text{and} \quad P(P = 1 \mid C = 1) = 0.9$$

- Moreover, knowing about fever no longer increases the knowledge about the test results and vice versa:

$$P(F = 1 \mid C = 1, P = 0) = P(F = 1 \mid C = 1, P = 1) = P(F = 1 \mid C = 1) = 0.6$$
$$P(P = 1 \mid C = 1, F = 0) = P(P = 1 \mid C = 1, F = 1) = P(P = 1 \mid C = 1) = 0.9$$

Conditional independence: The COVID example

Consider Fever and PCR Test Result.

- If someone has fever, we can guess their test result and vice versa:

$$P(F = 1 \mid P = 0) = 0.2 \quad \text{and} \quad P(F = 1 \mid P = 1) = 0.7$$
$$P(P = 1 \mid F = 0) = 0.3 \quad \text{and} \quad P(P = 1 \mid F = 1) = 0.4$$

$$\Longrightarrow \text{Fever and PCR are } \textbf{not} \text{ independent.}$$
$$F \not\perp P$$

- If we know the person has COVID ($C = 1$), they have fever and test positive with a high probability:

$$P(F = 1 \mid C = 1) = 0.6 \quad \text{and} \quad P(P = 1 \mid C = 1) = 0.9$$

- Moreover, knowing about fever no longer increases the knowledge about the test results and vice versa:

$$P(F = 1 \mid C = 1, P = 0) = P(F = 1 \mid C = 1, P = 1) = P(F = 1 \mid C = 1) = 0.6$$
$$P(P = 1 \mid C = 1, F = 0) = P(P = 1 \mid C = 1, F = 1) = P(P = 1 \mid C = 1) = 0.9$$
$$F \perp P \mid C = 1$$

Conditional Independence: The COVID example

- If we know the person does not have COVID ($C = 0$), they have fever and test positive with a low probability:

$$P(F = 1 \mid C = 0) = 0.2 \quad \text{and} \quad P(P = 1 \mid C = 0) = 0.1$$

Conditional Independence: The COVID example

- If we know the person does not have COVID ($C = 0$), they have fever and test positive with a low probability:

$$P(F = 1 \mid C = 0) = 0.2 \quad \text{and} \quad P(P = 1 \mid C = 0) = 0.1$$

- Also, knowing about the test result, does not change our knowledge about Fever and vice versa:

$$P(F = 1 \mid C = 0, P = 0) = P(F = 1 \mid C = 0, P = 1) = P(F = 1 \mid C = 0) = 0.2$$
$$P(P = 1 \mid C = 0, F = 0) = P(P = 1 \mid C = 0, F = 1) = P(P = 1 \mid C = 0) = 0.1$$

Conditional Independence: The COVID example

- If we know the person does not have COVID ($C = 0$), they have fever and test positive with a low probability:

$$P(F = 1 \mid C = 0) = 0.2 \quad \text{and} \quad P(P = 1 \mid C = 0) = 0.1$$

- Also, knowing about the test result, does not change our knowledge about Fever and vice versa:

$$P(F = 1 \mid C = 0, P = 0) = P(F = 1 \mid C = 0, P = 1) = P(F = 1 \mid C = 0) = 0.2$$
$$P(P = 1 \mid C = 0, F = 0) = P(P = 1 \mid C = 0, F = 1) = P(P = 1 \mid C = 0) = 0.1$$

$$\Rightarrow F \perp P \mid C = 0$$

$\Longrightarrow$ Fever is conditionally independent of the test result given COVID:

$$F \perp P \mid C.$$

Conditional Independence

Consider the sets of random variables $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$.

Definition

We say that $\boldsymbol{X}$ is *conditionally independent* of $\boldsymbol{Y}$ given $\boldsymbol{Z}$ in a distribution P, denoted $P \models (\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$, if

$$P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{Z} = \boldsymbol{z}) = P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{Z} = \boldsymbol{z}) \qquad \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}.$$

Conditional Independence

Consider the sets of random variables $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$.

Definition

We say that $\boldsymbol{X}$ is *conditionally independent* of $\boldsymbol{Y}$ given $\boldsymbol{Z}$ in a distribution P, denoted $P \models (\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$, if

$$P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{Z} = \boldsymbol{z}) = P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{Z} = \boldsymbol{z}) \qquad \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}.$$

- The variables in the set $\boldsymbol{Z}$ are said to be *observed*.
- If $\boldsymbol{Z}$ is empty, we write $(\boldsymbol{X} \perp \boldsymbol{Y})$ and say that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are marginally independent and $P(\boldsymbol{X} \mid \boldsymbol{Y}) = P(\boldsymbol{X})$.
- The set of all probability independencies in P is denoted by $\mathcal{I}(P)$.

Proposition

$P \models (\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$ if and only if

$$P(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z}) = P(\boldsymbol{X} \mid \boldsymbol{Z}) P(\boldsymbol{Y} \mid \boldsymbol{Z}).$$

How can conditional independence help?

- Back to the **C**OVID, **F**ever, **P**CR Test Result example, using **conditional probability**,

$$P(F, P, C) = P(F \mid P, C)P(P, C).$$

How can conditional independence help?

- Back to the **C**OVID, **F**ever, **P**CR Test Result example, using **conditional probability**,

$$P(F, P, C) = P(F \mid P, C)P(P, C).$$

- If $(F \perp P \mid C) \in \mathcal{I}(P)$, i.e., fever and test result are independent conditioned on COVID, then

$$P(F, P, C) = P(F \mid C)P(P, C).$$

How can conditional independence help?

- Back to the **C**OVID, **F**ever, **P**CR Test Result example, using **conditional probability**,

$$P(F, P, C) = P(F \mid P, C)P(P, C).$$

- If $(F \perp P \mid C) \in \mathcal{I}(P)$, i.e., fever and test result are independent conditioned on COVID, then

$$P(F, P, C) = P(F \mid C)P(P, C).$$

- Number of parameters $= 2 + (2^2 - 1) = 5$.

The importance of conditional independence

Conditional independence can reduce the number of parameters.

Exercise: Show that the following factorization results in the same number of parameters:

$$P(F, P, C) = P(F, P \mid C)P(C).$$

Chain rule

- How to generalize the idea?
- First, we need to factorize the joint distribution into CPDs.

Chain rule for random variables

$$P(X_1, \ldots, X_n) = P(X_n \mid X_{n-1}, \ldots, X_1) \ldots P(X_3 \mid X_2, X_1) P(X_2 \mid X_1) P(X_1)$$

$$= \prod_{i=1}^{n} P(X_i \mid X_{i-1}, \ldots, X_1)$$

Chain rule

- How to generalize the idea?
- First, we need to factorize the joint distribution into CPDs.

Chain rule for random variables

$$P(X_1, \ldots, X_n) = P(X_n \mid X_{n-1}, \ldots, X_1) \ldots P(X_3 \mid X_2, X_1) P(X_2 \mid X_1) P(X_1)$$

$$= \prod_{i=1}^{n} P(X_i \mid X_{i-1}, \ldots, X_1)$$

$\rightarrow$ $P(X_1, \ldots, X_n)$ may be written as $P(X_n, \ldots, X_1)$ to resemble the order of the factorization.

Exercise: For binary-valued random variables, compute the number of parameters required for the factorized term provided by the chain rule. Does chain rule reduce the number of parameters?

Hint: The number of parameters for $P(X_i \mid X_{i-1}, \cdots, X_1)$ is 2^{i-1}.

Conditional Independence + Chain Rule

- Consider joint distribution $P(X_1, X_2, X_3, X_4)$. Using the chain rule,

$$P(X_1, X_2, X_3, X_4) = P(X_4 \mid X_3, X_2, X_1)P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)$$

Conditional Independence + Chain Rule

- Consider joint distribution $P(X_1, X_2, X_3, X_4)$. Using the chain rule,

$$P(X_1, X_2, X_3, X_4) = P(X_4 \mid X_3, X_2, X_1)P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)$$

- If $(X_4 \perp X_1, X_2 \mid X_3) \in \mathcal{I}(P)$, then $P(X_4 \mid X_3, X_2, X_1) = P(X_4 \mid X_3)$.

Conditional Independence + Chain Rule

- Consider joint distribution $P(X_1, X_2, X_3, X_4)$. Using the chain rule,

$$P(X_1, X_2, X_3, X_4) = P(X_4 \mid X_3, X_2, X_1)P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)$$

- If $(X_4 \perp X_1, X_2 \mid X_3) \in \mathcal{I}(P)$, then $P(X_4 \mid X_3, X_2, X_1) = P(X_4 \mid X_3)$.
  - For binary variables, the number of required parameters for this CPD reduces from $2^3 = 8$ to $2^1 = 2$.
- The joint distribution becomes

$$P(X_1, X_2, X_3, X_4) = P(X_4 \mid X_3)P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)$$

  - The total number of parameters reduces from $2^4 - 1 = 15$ to $2 + 2^2 + 2 + 1 = 9$.

Conditional Independence + Chain Rule

- What if we additionally have $(X_1 \perp X_2 \mid X_3) \in \mathcal{I}(P)$?
- Then $P(X_1 \mid X_2, X_3) = P(X_1 \mid X_3)$. But we cannot apply this to the previous factorization:

$$P(X_1, X_2, X_3, X_4) = P(X_4 \mid X_3)P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)$$

Conditional Independence + Chain Rule

- What if we additionally have $(X_1 \perp X_2 \mid X_3) \in \mathcal{I}(P)$?
- Then $P(X_1 \mid X_2, X_3) = P(X_1 \mid X_3)$. But we cannot apply this to the previous factorization:

$$P(X_1, X_2, X_3, X_4) = P(X_4 \mid X_3)P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)$$

- So we need to re-factorize using the chain rule with a different ordering on the variables:

$$\begin{aligned} P(X_2, X_3, X_1, X_4) &= P(X_4 \mid X_1, X_3, X_2)P(X_1 \mid X_3, X_2)P(X_3 \mid X_2)P(X_2) \\ &= P(X_4 \mid X_3)P(X_1 \mid X_3)P(X_3 \mid X_2)P(X_2) \end{aligned}$$

- Now what if we additionally have $(X_3 \perp X_2 \mid X_1) \in \mathcal{I}(P)$?

Conditional Independence + Chain Rule

- What if we additionally have $(X_1 \perp X_2 \mid X_3) \in \mathcal{I}(P)$?
- Then $P(X_1 \mid X_2, X_3) = P(X_1 \mid X_3)$. But we cannot apply this to the previous factorization:

$$P(X_1, X_2, X_3, X_4) = P(X_4 \mid X_3)P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)$$

- So we need to re-factorize using the chain rule with a different ordering on the variables:

$$\begin{aligned} P(X_2, X_3, X_1, X_4) &= P(X_4 \mid X_1, X_3, X_2)P(X_1 \mid X_3, X_2)P(X_3 \mid X_2)P(X_2) \\ &= P(X_4 \mid X_3)P(X_1 \mid X_3)P(X_3 \mid X_2)P(X_2) \end{aligned}$$

- Now what if we additionally have $(X_3 \perp X_2 \mid X_1) \in \mathcal{I}(P)$?
- It is impossible to further simply the factorization using this independence.

$\Rightarrow$ Not all independencies may appear in a factorization.

See it in practice

Back to the COVID problem with 12 variables:

#	COVID	Mask	Social Distancing	FIU	COugh	Fever	Ventilation	Season	ConGestion	Difficulty Breathing	DRug	Allergy
1	True	False	True	False	True	True	True	Spring	True	True	False	False
2	False	True	True	False	False	True	False	Summer	False	False	True	False
3	True	False	False	True	True	False	False	Fall	False	True	True	False
4	False	True	True	False	False	True	False	Winter	True	True	False	True
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1000	True	True	True	False	True	False	True	Spring	False	False	True	True

So we just find the conditional independencies and factorize the joint distribution accordingly? Yes, but...

1. How to find the conditional independencies $\mathcal{I}(P)$ from the data?

See it in practice

Back to the COVID problem with 12 variables:

#	COVID	Mask	Social Distancing	FIU	COugh	Fever	Ventilation	Season	ConGestion	Difficulty Breathing	DRug	Allergy
1	True	False	True	False	True	True	True	Spring	True	True	False	False
2	False	True	True	False	False	True	False	Summer	False	False	True	False
3	True	False	False	True	True	False	False	Fall	False	True	True	False
4	False	True	True	False	False	True	False	Winter	True	True	False	True
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1000	True	True	True	False	True	False	True	Spring	False	False	True	True

So we just find the conditional independencies and factorize the joint distribution accordingly? Yes, but...

1. How to find the conditional independencies $\mathcal{I}(P)$ from the data? $\rightarrow$ Chapter: Structure learning (there are statistical tests for this)
2. Given the conditional independencies, how to factorize the joint distribution? (which ones to include and according to what order?)

Summary: What is the goal?

- The goal is to find from data, the **"correct" factorization** of the joint probability distribution.

$$P(C, M, D) = \begin{cases} P(C)P(M \mid C)P(D) & \text{4 parameters} \\ P(C \mid M)P(M)P(D) & \text{4 parameters} \\ \vdots & \\ P(C \mid M)P(M \mid D)P(D) & \text{5 parameters} \\ P(M \mid C, D)P(C)P(D) & \text{6 parameters} \end{cases}$$

#	C	M	D
1	True	False	True
2	False	True	True
3	True	False	False
4	False	True	True
5	False	True	True
6	False	True	False
7	True	False	True
8	False	True	True
9	False	True	True
10	False	False	False
11	False	True	False
12	True	True	True

Factorizing the joint distribution: Graphical visualization

- Back to the example with $(X_4 \perp X_1, X_2 \mid X_3) \in \mathcal{I}(P)$, resulting in

$$P(X_1, X_2, X_3, X_4) = P(X_4 \mid \underbrace{X_3}_{\mathrm{Pa}_{X_4}}) P(X_3 \mid \underbrace{X_1, X_2}_{\mathrm{Pa}_{X_3}}) P(X_2 \mid \underbrace{X_1}_{\mathrm{Pa}_{X_2}}) P(X_1)$$

Factorizing the joint distribution: Graphical visualization

- Back to the example with $(X_4 \perp X_1, X_2 \mid X_3) \in \mathcal{I}(P)$, resulting in

$$P(X_1, X_2, X_3, X_4) = P(X_4 \mid \underbrace{X_3}_{\mathrm{Pa}_{X_4}}) P(X_3 \mid \underbrace{X_1, X_2}_{\mathrm{Pa}_{X_3}}) P(X_2 \mid \underbrace{X_1}_{\mathrm{Pa}_{X_2}}) P(X_1)$$

- For each X_i, define the *parents of* X_i, denoted Pa_{X_i}, as the set of variables that X_i is conditioned on in the factorization:

$$\mathrm{Pa}_{X_4} = \{X_3\}, \quad \mathrm{Pa}_{X_3} = \{X_1, X_2\}, \quad \mathrm{Pa}_{X_2} = \{X_1\}, \quad \mathrm{Pa}_{X_1} = \emptyset$$

Factorizing the joint distribution: Graphical visualization

- Back to the example with $(X_4 \perp X_1, X_2 \mid X_3) \in \mathcal{I}(P)$, resulting in

$$P(X_1, X_2, X_3, X_4) = P(X_4 \mid \underbrace{X_3}_{\mathrm{Pa}_{X_4}}) P(X_3 \mid \underbrace{X_1, X_2}_{\mathrm{Pa}_{X_3}}) P(X_2 \mid \underbrace{X_1}_{\mathrm{Pa}_{X_2}}) P(X_1)$$

- For each X_i, define the *parents of* X_i, denoted Pa_{X_i}, as the set of variables that X_i is conditioned on in the factorization:

$$\mathrm{Pa}_{X_4} = \{X_3\}, \quad \mathrm{Pa}_{X_3} = \{X_1, X_2\}, \quad \mathrm{Pa}_{X_2} = \{X_1\}, \quad \mathrm{Pa}_{X_1} = \emptyset$$

- Then the joint distribution can be written as

$$P(X_1, X_2, X_3, X_4) = \prod_{i=1}^{4} P(X_i \mid \mathrm{Pa}_{X_i})$$

- So each node is conditioned (depends) on only its parents in the factorization.

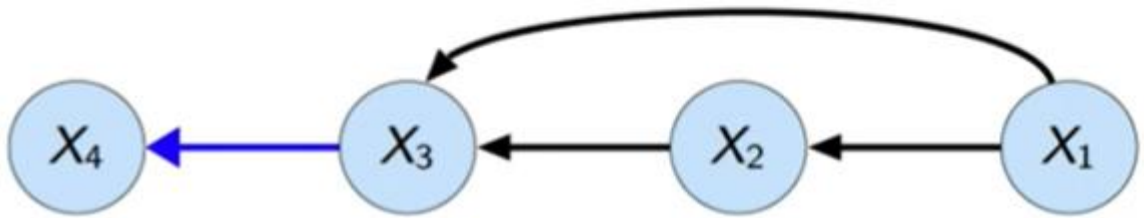
Factorizing the joint distribution: Graphical visualization

- Now construct the directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertices $\mathcal{V} = \{X_1, X_2, X_3, X_4\}$ and where $\mathcal{E}$ is the set of directed edges from Pa_{X_i} to X_i for $i = 1, 2, 3, 4$.

Factorizing the joint distribution: Graphical visualization

- Now construct the directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertices $\mathcal{V} = \{X_1, X_2, X_3, X_4\}$ and where $\mathcal{E}$ is the set of directed edges from Pa_{X_i} to X_i for $i = 1, 2, 3, 4$.

Directed Acyclic Graph (DAG):

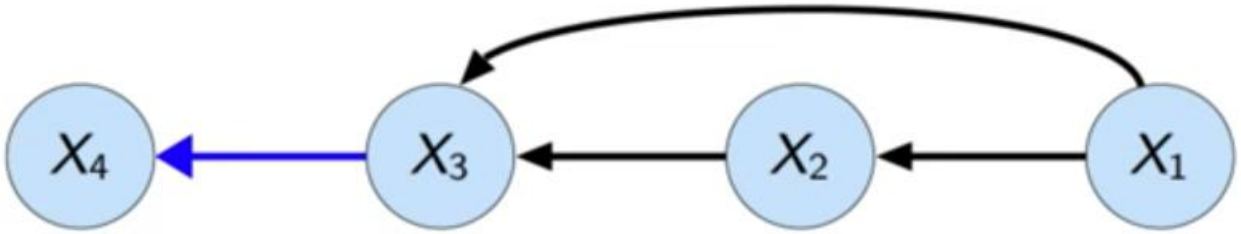


Conditional Probability Distributions (CPDs):

$P(X_4 \mid X_3)$ $P(X_3 \mid X_2, X_1)$ $P(X_2 \mid X_1)$ $P(X_1)$

- Now construct the directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertices $\mathcal{V} = \{X_1, X_2, X_3, X_4\}$ and where $\mathcal{E}$ is the set of directed edges from Pa_{X_i} to X_i for $i = 1, 2, 3, 4$.

Directed Acyclic Graph (DAG):



Conditional Probability Distributions (CPDs):

$P(X_4 \mid X_3)$ $P(X_3 \mid X_2, X_1)$ $P(X_2 \mid X_1)$ $P(X_1)$

Bayesian Network

Joint Distribution:

$$P(X_1, X_2, X_3, X_4) = P(X_4 \mid X_3)P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)$$

Exercise: Show that the graph is always acyclic (does not have a directed cycle).

Hint: The graph is based on the chain-rule factorization.

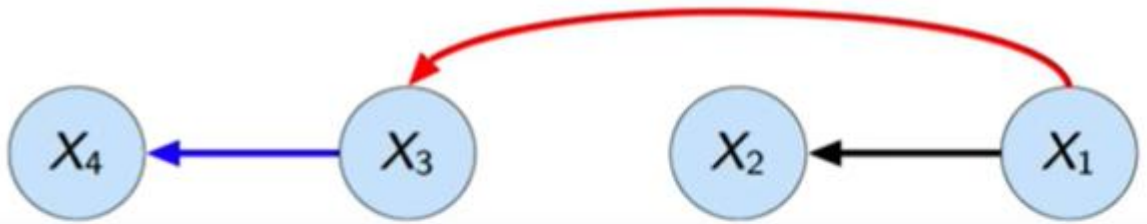
- X_1, X_2, X_3, X_4 is called a *topological ordering relative to the DAG* $\mathcal{G}$ as X_i is connected to X_j only if $i < j$.

Factorizing the joint distribution: Graphical visualization

- If we have both $(X_4 \perp X_1, X_2 \mid X_3), (X_3 \perp X_2 \mid X_1) \in \mathcal{I}(P)$, then

$$\begin{aligned} P(X_1, X_2, X_3, X_4) &= P(X_4 \mid X_3, X_2, X_1) P(X_3 \mid X_2, X_1) P(X_2 \mid X_1) P(X_1) \\ &= P(X_4 \mid X_3) P(X_3 \mid X_1) P(X_2 \mid X_1) P(X_1) \end{aligned}$$

Directed Acyclic Graph (DAG):



Conditional Probability Distributions (CPDs):

$P(X_4 \mid X_3)$ $P(X_3 \mid X_1)$ $P(X_2 \mid X_1)$ $P(X_1)$

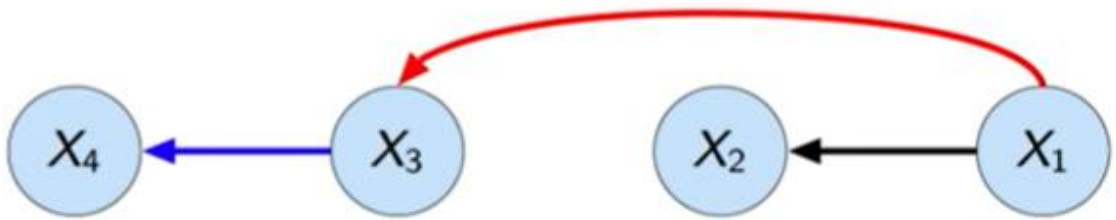
Bayesian Network

Factorizing the joint distribution: Graphical visualization

- If we have both $(X_4 \perp X_1, X_2 \mid X_3), (X_3 \perp X_2 \mid X_1) \in \mathcal{I}(P)$, then

$$\begin{aligned} P(X_1, X_2, X_3, X_4) &= P(X_4 \mid X_3, X_2, X_1) P(X_3 \mid X_2, X_1) P(X_2 \mid X_1) P(X_1) \\ &= P(X_4 \mid X_3) P(X_3 \mid X_1) P(X_2 \mid X_1) P(X_1) \end{aligned}$$

Directed Acyclic Graph (DAG):



Conditional Probability Distributions (CPDs):

$$P(X_4 \mid \mathrm{Pa}_{X_4}) P(X_3 \mid \mathrm{Pa}_{X_3}) P(X_2 \mid \mathrm{Pa}_{X_2}) P(X_1 \mid \mathrm{Pa}_{X_1})$$

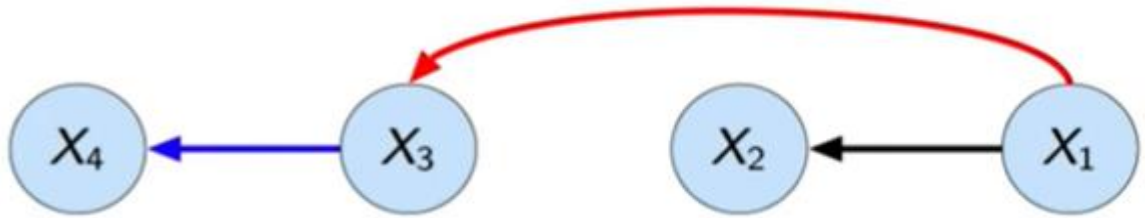
Bayesian Network

Factorizing the joint distribution: Graphical visualization

- If we have both $(X_4 \perp X_1, X_2 \mid X_3), (X_3 \perp X_2 \mid X_1) \in \mathcal{I}(P)$, then

$$\begin{aligned} P(X_1, X_2, X_3, X_4) &= P(X_4 \mid X_3, X_2, X_1) P(X_3 \mid X_2, X_1) P(X_2 \mid X_1) P(X_1) \\ &= P(X_4 \mid X_3) P(X_3 \mid X_1) P(X_2 \mid X_1) P(X_1) \end{aligned}$$

Directed Acyclic Graph (DAG):



Conditional Probability Distributions (CPDs):

$$P(X_4 \mid \mathrm{Pa}_{X_4}) P(X_3 \mid \mathrm{Pa}_{X_3}) P(X_2 \mid \mathrm{Pa}_{X_2}) P(X_1 \mid \mathrm{Pa}_{X_1})$$

Bayesian Network

Joint Distribution:

$$\prod_{i=1}^{4} P(X_i \mid \mathrm{Pa}_{X_i})$$

Bayesian network

Definition: Factorization (Chain rule for Bayesian networks)

The distribution P **factorizes** according to the DAG $\mathcal{G}$, if

$$P(X_1, \cdots, X_n) = \prod_{i=1}^{n} P(X_i \mid \mathrm{Pa}_{X_i}^{\mathcal{G}}).$$

Definition: Factorization (Chain rule for Bayesian networks)

The distribution P **factorizes** according to the DAG $\mathcal{G}$, if

$$P(X_1, \cdots, X_n) = \prod_{i=1}^{n} P(X_i \mid \mathrm{Pa}_{X_i}^{\mathcal{G}}).$$

Note: $\mathrm{Pa}_{X_i}^{\mathcal{G}}$ denotes the parents of X_i in graph $\mathcal{G}$ which was equal to the chain-rule-based definition Pa_{X_i} in the previous examples.

Definition:Bayesian Network

Given the random variables $\mathcal{V} = \{X_1, \ldots, X_n\}$, a *Bayesian Network (BN)* is a pair $\mathcal{B} = (\mathcal{G}; P_{\mathcal{B}})$, where

- $\mathcal{G}$ is a directed acyclic graph (**BN structure**) with node set $\mathcal{V}$,
- $P_{\mathcal{B}}$ is a probability function that factorizes according to $\mathcal{G}$ and is specified as a set of conditional probability distributions **(CPD)**s $P_{\mathcal{B}}(X_i \mid \mathrm{Pa}_{X_i})$ for all $X_i \in \mathcal{V}$ (**BN parameters**).

See it in practice: Obtaining the Bayesian network

- Assume the following joint probability distribution for **M**ask, Social **D**istancing, **C**OVID, **F**ever, Difficulty **B**reathing, and **V**entilation:

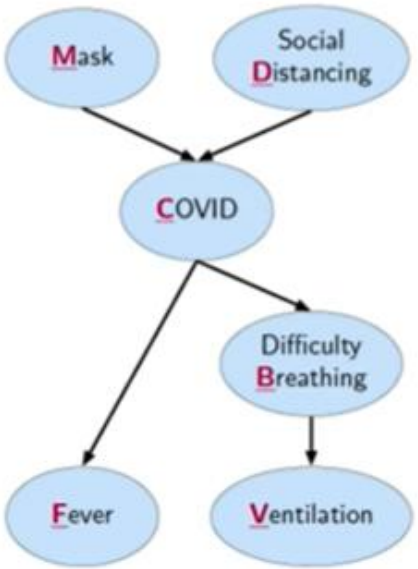
$$P(M, D, C, B, F, V) = P(F \mid C)P(V \mid B)P(B \mid C)P(C \mid M, D)P(M)P(D)$$

See it in practice: Obtaining the Bayesian network

- Assume the following joint probability distribution for **M**ask, Social **D**istancing, **C**OVID, **F**ever, Difficulty **B**reathing, and **V**entilation:

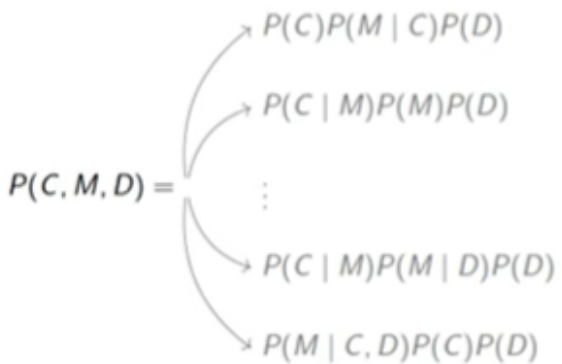
$$P(M, D, C, B, F, V) = P(F \mid C)P(V \mid B)P(B \mid C)P(C \mid M, D)P(M)P(D)$$

- Obtain the corresponding Bayesian network.



Summary: What is the goal?

- The goal is to find from data, the **"correct" factorization** of the joint probability distribution.
  - How? By finding the conditional independencies $\mathcal{I}(P)$.
  - And then?



$\mathcal{I}(P) =$
$\{(C \perp D \mid M), (M \perp D)\}$

#	C	M	D
1	True	False	True
2	False	True	True
3	True	False	False
4	False	True	True
5	False	True	True
6	False	True	False
7	True	False	True
8	False	True	True
9	False	True	True
10	False	False	False
11	False	True	False
12	True	True	True

From factorization to independence

- Given a factorization, how to obtain the conditional independencies?

$$P(X_1, X_2, X_3, X_4) = P(X_4 \mid X_3)P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1) \quad (*)$$

From factorization to independence

- Given a factorization, how to obtain the conditional independencies?

$$P(X_1, X_2, X_3, X_4) = P(X_4 \mid X_3)P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1) \quad (*)$$

- Compare it to its original chain-rule factorization:

$$P(X_1, X_2, X_3, X_4) = P(X_4 \mid X_3, X_2, X_1)P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)$$

From factorization to independence

- Given a factorization, how to obtain the conditional independencies?

$$P(X_1, X_2, X_3, X_4) = P(X_4 \mid X_3)P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1) \quad (*)$$

- Compare it to its original chain-rule factorization:

$$P(X_1, X_2, X_3, X_4) = P(X_4 \mid X_3, X_2, X_1)P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)$$

- Only X_3 is the parent of X_4. Hence, X_4 should become independent from the other variables that it depends on in the chain rule factorization, i.e., X_1 and X_2, conditioned on its parent X_3:

$$(X_4 \perp X_2, X_1 \mid X_3) \in \mathcal{I}(P).$$

From factorization to independence

- Is it also necessary? That is, does the following factorization

$$P(X_1, X_2, X_3, X_4) = P(X_4 \mid X_3)P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1) \quad (*)$$

imply the following conditional independence?

$$(X_4 \perp X_2, X_1 \mid X_3) \in \mathcal{I}(P).$$

- We have

$$\begin{aligned} P(X_4 \mid X_3, X_2, X_1) &= \frac{P(X_4, X_3, X_2, X_1)}{P(X_3, X_2, X_1)} \\ &= \frac{P(X_4 \mid X_3)P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)}{\sum_{X_4} P(X_4 \mid X_3)P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)} \\ &= \frac{P(X_4 \mid X_3)}{\sum_{X_4} P(X_4 \mid X_3)} \\ &= P(X_4 \mid X_3) \end{aligned}$$

From factorization to independence

- What if the original chain-rule-based factorization was not so clear?

$$P(X_1, X_2, X_3, X_4) = P(X_4 \mid X_1)P(X_1 \mid X_3)P(X_2 \mid X_4)P(X_3) \qquad (*)$$

From factorization to independence

- What if the original chain-rule-based factorization was not so clear?

$$P(X_1, X_2, X_3, X_4) = P(X_4 \mid X_1)P(X_1 \mid X_3)P(X_2 \mid X_4)P(X_3) \qquad (*)$$

- To find the "correct" ordering, note that in the chain rule,
  - Once a variable X_i appears on the left of the conditioning, it no longer appears in the remaining CPDs on its right:

$$P(X_1, \ldots, X_n) = P(X_n \mid X_{n-1}, \ldots, X_1) \ldots P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)$$

From factorization to independence

- What if the original chain-rule-based factorization was not so clear?

$$P(X_1, X_2, X_3, X_4) = P(X_4 \mid X_1)P(X_1 \mid X_3)P(X_2 \mid X_4)P(X_3) \qquad (*)$$

- To find the "correct" ordering, note that in the chain rule,
  - Once a variable X_i appears on the left of the conditioning, it no longer appears in the remaining CPDs on its right:

$$P(X_1, \ldots, X_n) = P(X_n \mid X_{n-1}, \ldots, X_1) \ldots P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)$$

- So we can re-order $(*)$ as

$$P(X_1, X_2, X_3, X_4) = P(X_2 \mid X_4)P(X_4 \mid X_1)P(X_1 \mid X_3)P(X_3)$$

From factorization to independence

- What if the original chain-rule-based factorization was not so clear?

$$P(X_1, X_2, X_3, X_4) = P(X_4 \mid X_1)P(X_1 \mid X_3)P(X_2 \mid X_4)P(X_3) \qquad (*)$$

- To find the "correct" ordering, note that in the chain rule,
  - Once a variable X_i appears on the left of the conditioning, it no longer appears in the remaining CPDs on its right:

$$P(X_1, \ldots, X_n) = P(X_n \mid X_{n-1}, \ldots, X_1) \ldots P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)$$

- So we can re-order $(*)$ as

$$P(X_1, X_2, X_3, X_4) = P(X_2 \mid X_4)P(X_4 \mid X_1)P(X_1 \mid X_3)P(X_3)$$

- and the original chain-rule-based factorization

$$P(X_3, X_1, X_4, X_2) = P(X_2 \mid X_4, X_1, X_3)P(X_4 \mid X_1, X_3)P(X_1 \mid X_3)P(X_3)$$

From factorization to independence

- What if the original chain-rule-based factorization was not so clear?

$$P(X_1, X_2, X_3, X_4) = P(X_4 \mid X_1)P(X_1 \mid X_3)P(X_2 \mid X_4)P(X_3) \qquad (*)$$

- To find the "correct" ordering, note that in the chain rule,
  - Once a variable X_i appears on the left of the conditioning, it no longer appears in the remaining CPDs on its right:

$$P(X_1, \ldots, X_n) = P(X_n \mid X_{n-1}, \ldots, X_1) \ldots P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)$$

- So we can re-order $(*)$ as

$$P(X_1, X_2, X_3, X_4) = P(X_2 \mid X_4)P(X_4 \mid X_1)P(X_1 \mid X_3)P(X_3)$$

- and the original chain-rule-based factorization

$$P(X_3, X_1, X_4, X_2) = P(X_2 \mid X_4, X_1, X_3)P(X_4 \mid X_1, X_3)P(X_1 \mid X_3)P(X_3)$$

$$\Rightarrow (X_2 \perp X_1, X_3 \mid X_4), (X_4 \perp X_3 \mid X_1) \in \mathcal{I}(P).$$

From factorization to independence: Using BNs

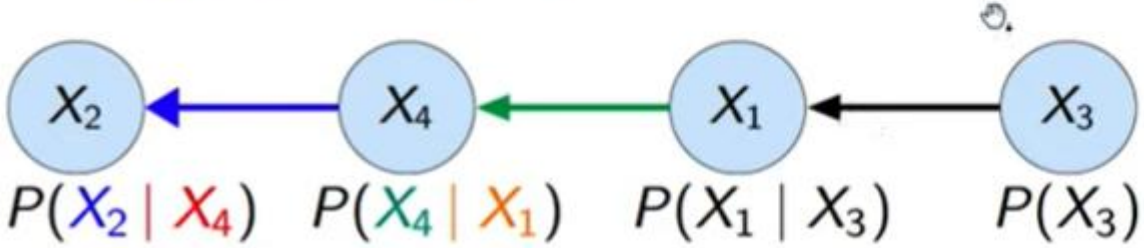
- What is the equivalent graphical condition?

$$P(X_1, X_2, X_3, X_4) = P(X_2 \mid X_4)P(X_4 \mid X_1)P(X_1 \mid X_3)P(X_3)$$

From factorization to independence: Using BNs

- What is the equivalent graphical condition?

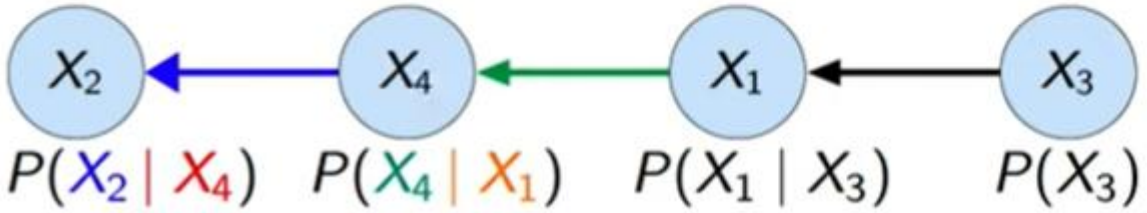
$$P(X_1, X_2, X_3, X_4) = P(X_2 \mid X_4)P(X_4 \mid X_1)P(X_1 \mid X_3)P(X_3)$$



From factorization to independence: Using BNs

- What is the equivalent graphical condition?

$$P(X_1, X_2, X_3, X_4) = P(X_2 \mid X_4)P(X_4 \mid X_1)P(X_1 \mid X_3)P(X_3)$$



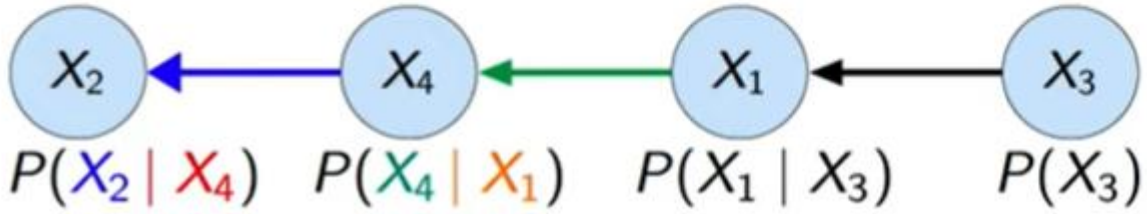
- Then

$$(X_2 \perp X_1, X_3 \mid X_4), (X_4 \perp X_3 \mid X_1) \in \mathcal{I}(P)$$

From factorization to independence: Using BNs

- What is the equivalent graphical condition?

$$P(X_1, X_2, X_3, X_4) = P(X_2 \mid X_4)P(X_4 \mid X_1)P(X_1 \mid X_3)P(X_3)$$



- Then

$$(X_2 \perp X_1, X_3 \mid X_4), (X_4 \perp X_3 \mid X_1) \in \mathcal{I}(P)$$

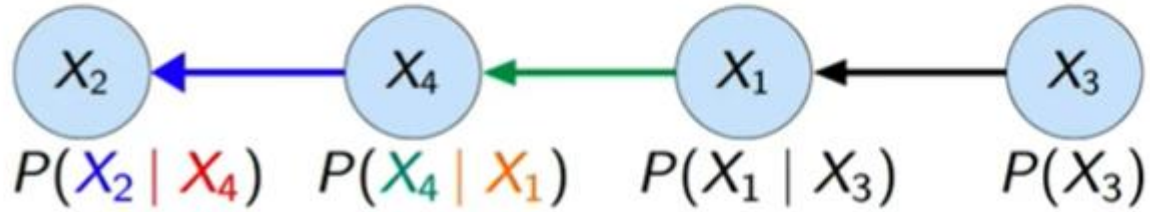
$$\Updownarrow$$

$$(X_2 \perp \text{NonDescendants}_{X_2} \mid \text{Pa}_{X_2}), (X_4 \perp \text{NonDescendants}_{X_4} \mid \text{Pa}_{X_4}) \in \mathcal{I}(P).$$

From factorization to independence: Using BNs

- What is the equivalent graphical condition?

$$P(X_1, X_2, X_3, X_4) = P(X_2 \mid X_4)P(X_4 \mid X_1)P(X_1 \mid X_3)P(X_3)$$



- Then

$$(X_2 \perp X_1, X_3 \mid X_4), (X_4 \perp X_3 \mid X_1) \in \mathcal{I}(P)$$

$$\Updownarrow$$

$$(X_2 \perp \text{NonDescendants}_{X_2} \mid \text{Pa}_{X_2}), (X_4 \perp \text{NonDescendants}_{X_4} \mid \text{Pa}_{X_4}) \in \mathcal{I}(P).$$

- where $\text{NonDescendants}_{X_i}$ are all the nodes excluding the *descendants* of X_i (to which X_i is connected by a directed path).
- These conditional independencies are known as the *local independencies* of the graph, because they are conditioned on X_i's parents (that are local to X_i).

$\mathcal{I}_l(\mathcal{G})$ and $\mathcal{I}(P)$

Definition: Local (Markov) independence

Given the graph $\mathcal{G}$, the set of **local independencies**, denoted by $\mathcal{I}_l(\mathcal{G})$, consists of

$$(X_i \perp \text{NonDescendants}_{X_i} \mid \text{Pa}_{X_i}^{\mathcal{G}}) \qquad \forall i.$$

$\mathcal{I}_l(\mathcal{G})$ and $\mathcal{I}(P)$

Definition: Local (Markov) independence

Given the graph $\mathcal{G}$, the set of **local independencies**, denoted by $\mathcal{I}_l(\mathcal{G})$, consists of

$$(X_i \perp \text{NonDescendants}_{X_i} \mid \text{Pa}_{X_i}^{\mathcal{G}}) \qquad \forall i.$$

Definition: Independence-map

$\mathcal{G}$ is an I-map for P if $\mathcal{I}_l(\mathcal{G}) \subseteq \mathcal{I}(P)$.

- So $\mathcal{G}$ being an I-map for P means that P satisfies the local independencies of $\mathcal{G}$.

$\mathcal{I}_l(\mathcal{G})$ and $\mathcal{I}(P)$

Definition: Local (Markov) independence

Given the graph $\mathcal{G}$, the set of **local independencies**, denoted by $\mathcal{I}_l(\mathcal{G})$, consists of

$$(X_i \perp \text{NonDescendants}_{X_i} \mid \text{Pa}^{\mathcal{G}}_{X_i}) \qquad \forall i.$$

Definition: Independence-map

$\mathcal{G}$ is an I-map for P if $\mathcal{I}_l(\mathcal{G}) \subseteq \mathcal{I}(P)$.

- So $\mathcal{G}$ being an I-map for P means that P satisfies the local independencies of $\mathcal{G}$.

(The I-map) Theorem

Let $\mathcal{G}$ be a DAG and P be a joint distribution over a set of random variables.

- P factorizes according to $\mathcal{G}$, if and only if $\mathcal{G}$ is an I-map for P.

$\mathcal{I}_l(\mathcal{G})$ and $\mathcal{I}(P)$: Example

- Consider the joint distribution P over $X_1, \ldots, X_4$, where

$$\mathcal{I}(P) = \{(X_4 \perp X_2, X_1 \mid X_3) \text{ and its derivations}\}.$$

Recall: Probabilistic independence rules

- $(X_4 \perp X_2, X_1 | X_3)$ is indeed the same as $(X_4 \perp (X_2, X_1) | X_3)$.
- Although the separate independencies are concluded:

$$(X_4 \perp X_2, X_1 | X_3) \Rightarrow (X_4 \perp X_2 | X_3), (X_4 \perp X_1 | X_3),$$

Recall: Probabilistic independence rules

- $(X_4 \perp X_2, X_1 | X_3)$ is indeed the same as $(X_4 \perp (X_2, X_1) | X_3)$.
- Although the separate independencies are concluded:

$$(X_4 \perp X_2, X_1 | X_3) \Rightarrow (X_4 \perp X_2 | X_3), (X_4 \perp X_1 | X_3),$$

the reverse does not hold:

$$(X_4 \perp X_2 | X_3), (X_4 \perp X_1 | X_3) \not\Rightarrow (X_4 \perp X_2, X_1 | X_3).$$

- The "conditioned" part can appear on the left as well, e.g.,

$$(X_4 \perp X_2 | X_3) \Rightarrow (X_4 \perp X_2, X_3 | X_3).$$

- The independent term can be "conditioned on," e.g.,

$$(X_4 \perp X_2 | X_3) \Rightarrow (X_4 \perp X_2 | X_3, X_2).$$

$\mathcal{I}_l(\mathcal{G})$ and $\mathcal{I}(P)$: Example

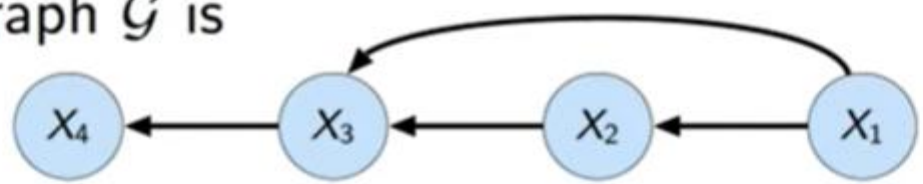
- Consider the joint distribution P over $X_1, \ldots, X_4$, where

$$\mathcal{I}(P) = \{(X_4 \perp X_2, X_1 \mid X_3) \text{ and its derivations}\}.$$

- Does P satisfy the following factorization?

$$P(X_1, \ldots, X_4) = P(X_4 \mid X_3)P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)$$

- The equivalent graph $\mathcal{G}$ is



imposing the local independencies

$$\mathcal{I}_l(\mathcal{G}) = \{(X_4 \perp X_2, X_1 \mid X_3)\}$$

$\mathcal{I}_l(\mathcal{G})$ and $\mathcal{I}(P)$: Example

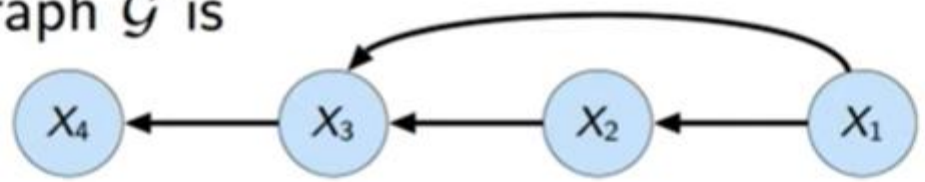
- Consider the joint distribution P over $X_1, \ldots, X_4$, where

$$\mathcal{I}(P) = \{(X_4 \perp X_2, X_1 \mid X_3) \text{ and its derivations}\}.$$

- Does P satisfy the following factorization?

$$P(X_1, \ldots, X_4) = P(X_4 \mid X_3)P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)$$

- The equivalent graph $\mathcal{G}$ is



imposing the local independencies

$$\mathcal{I}_l(\mathcal{G}) = \{(X_4 \perp X_2, X_1 \mid X_3)\}$$

$$\Rightarrow \mathcal{I}_l(\mathcal{G}) \subseteq \mathcal{I}(P) \qquad \mathcal{G} \text{ is an I-map for } P.$$

- So P satisfies the above factorization.

$\mathcal{I}_l(\mathcal{G})$ and $\mathcal{I}(P)$: Example

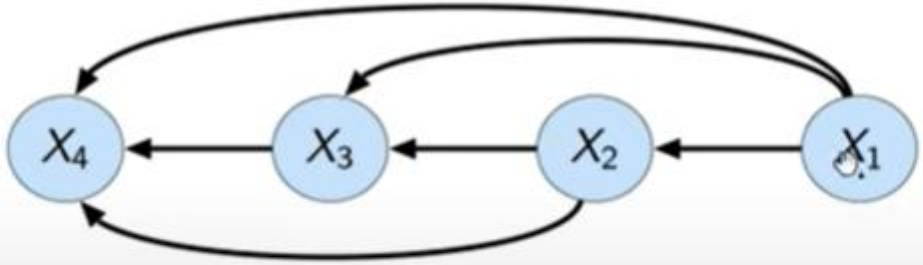
- Consider the joint distribution P over $X_1, \ldots, X_4$, where

$$\mathcal{I}(P) = \{(X_4 \perp X_2, X_1 \mid X_3) \text{ and its derivations}\}.$$

- Does P satisfy the following factorization?

$$P(X_1, \ldots, X_4) = P(X_4 \mid X_3, X_2, X_1)P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)$$

- The equivalent graph $\mathcal{G}$ is



imposing the local independencies

$$\mathcal{I}_l(\mathcal{G}) = \emptyset$$

$\mathcal{I}_l(\mathcal{G})$ and $\mathcal{I}(P)$: Example

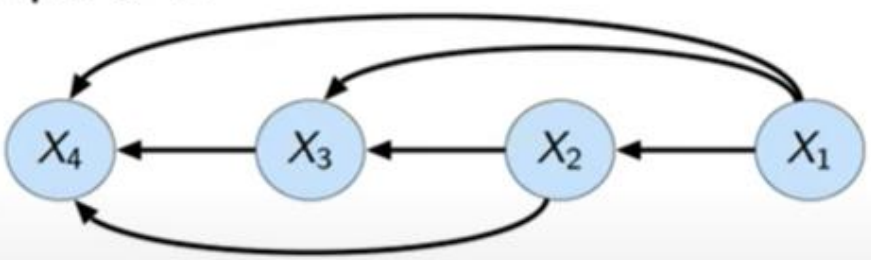
- Consider the joint distribution P over $X_1, \ldots, X_4$, where

$$\mathcal{I}(P) = \{(X_4 \perp X_2, X_1 \mid X_3) \text{ and its derivations}\}.$$

- Does P satisfy the following factorization?

$$P(X_1, \ldots, X_4) = P(X_4 \mid X_3, X_2, X_1)P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)$$

- The equivalent graph $\mathcal{G}$ is



imposing the local independencies

$$\mathcal{I}_l(\mathcal{G}) = \emptyset$$

$$\Rightarrow \mathcal{I}_l(\mathcal{G}) \subset \mathcal{I}(P) \qquad \mathcal{G} \text{ is an I-map for } P.$$

- So P does satisfy the above factorization.

$\mathcal{I}_l(\mathcal{G})$ and $\mathcal{I}(P)$: Example

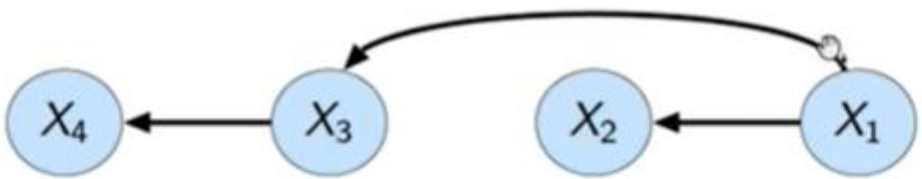
- Consider the joint distribution P over $X_1, \ldots, X_4$, where

$$\mathcal{I}(P) = \{(X_4 \perp X_2, X_1 \mid X_3) \text{ and its derivations}\}. \qquad (*)$$

- Does P satisfy the following factorization?

$$P(X_1, \ldots, X_4) = P(X_4 \mid X_3)P(X_3 \mid X_1)P(X_2 \mid X_1)P(X_1) \qquad (**)$$

- The equivalent graph $\mathcal{G}$ is



imposing the local independencies

$$\mathcal{I}_l(\mathcal{G}) = \{(X_4 \perp X_2, X_1 \mid X_3), (X_3 \perp X_2 \mid X_1), (X_2 \perp X_4, X_3 \mid X_1)\}$$

$$\Rightarrow \mathcal{I}_l(\mathcal{G}) \not\subseteq \mathcal{I}(P) \qquad \mathcal{G} \text{ is } \textbf{not} \text{ an I-map for } P.$$

- So P does not satisfy the above factorization.

$\mathcal{I}_l(\mathcal{G})$ and $\mathcal{I}(P)$: Example

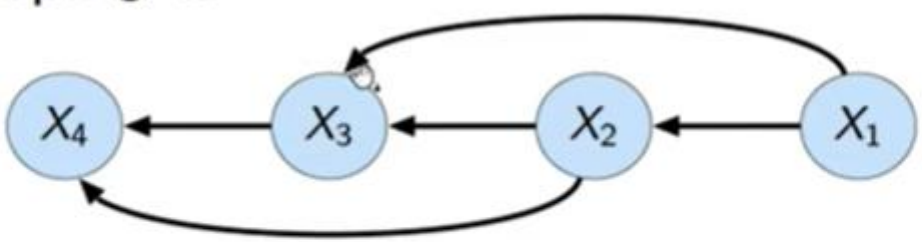
- Consider the joint distribution P over $X_1, \ldots, X_4$, where

$$\mathcal{I}(P) = \{(X_4 \perp X_2, X_1 \mid X_3) \text{ and its derivations}\}.$$

- Does P satisfy the following factorization?

$$P(X_1, \ldots, X_4) = P(X_4 \mid X_3, X_2)P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)$$

- The equivalent graph $\mathcal{G}$ is



$\mathcal{I}_l(\mathcal{G})$ and $\mathcal{I}(P)$: Example

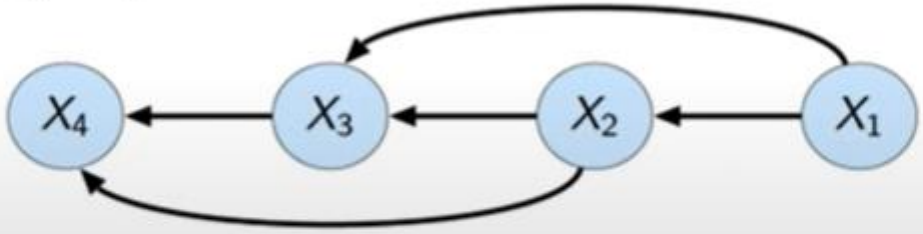
- Consider the joint distribution P over $X_1, \ldots, X_4$, where

$$\mathcal{I}(P) = \{(X_4 \perp X_2, X_1 \mid X_3) \text{ and its derivations}\}.$$

- Does P satisfy the following factorization?

$$P(X_1, \ldots, X_4) = P(X_4 \mid X_3, X_2)P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)$$

- The equivalent graph $\mathcal{G}$ is



imposing the local independencies

$$\mathcal{I}_l(\mathcal{G}) = \{(X_4 \perp X_1 \mid X_3, X_2)\}.$$

Minimal I-map

- Given a conditional independence, the distribution P does not factorize uniquely.
- Consider the distribution P defined on $X_1, \ldots, X_4$, where

$$\mathcal{I}(P) = \{(X_4 \perp X_1, X_2 \mid X_3) \text{ and its derivations}\}.$$

Minimal I-map

- Given a conditional independence, the distribution P does not factorize uniquely.
- Consider the distribution P defined on $X_1, \ldots, X_4$, where

$$\mathcal{I}(P) = \{(X_4 \perp X_1, X_2 \mid X_3) \text{ and its derivations}\}.$$

- Consider the following graphs:

$$\mathcal{I}_l(\mathcal{G}_1) = \{(X_4 \perp X_1, X_2 \mid X_3)\}$$

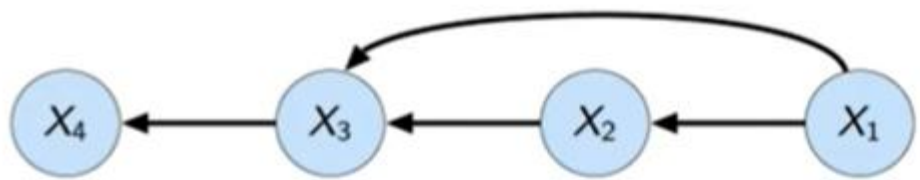


Figure: $\mathcal{G}_1$

Minimal I-map

- Given a conditional independence, the distribution P does not factorize uniquely.
- Consider the distribution P defined on $X_1, \ldots, X_4$, where

$$\mathcal{I}(P) = \{(X_4 \perp X_1, X_2 \mid X_3) \text{ and its derivations}\}.$$

- Consider the following graphs:

$$\mathcal{I}_l(\mathcal{G}_1) = \{(X_4 \perp X_1, X_2 \mid X_3)\}$$

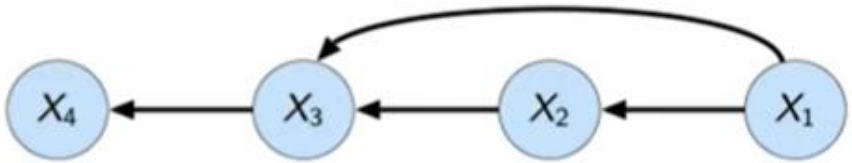


Figure: $\mathcal{G}_1$

$$\mathcal{I}_l(\mathcal{G}_2) = \emptyset$$

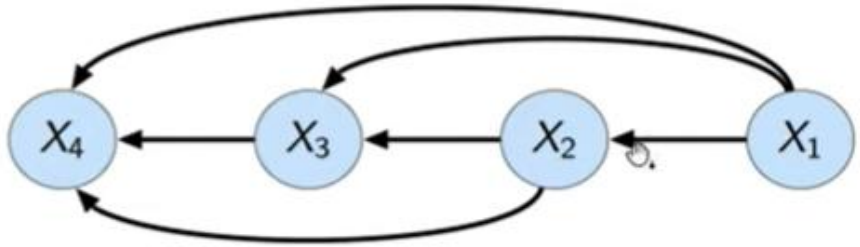


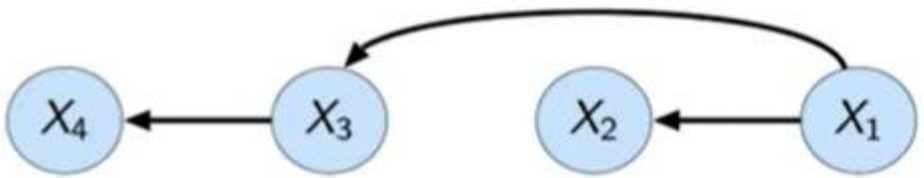
Figure: $\mathcal{G}_2$

$\Rightarrow$ Both $\mathcal{G}_1$ and $\mathcal{G}_2$ are an I-map for P

Minimal I-map

- Note that if the edges $X_2 \to X_4$ and $X_1 \to X_4$ are omitted from $\mathcal{G}_2$, we obtain $\mathcal{G}_1$, which is still an I-map for P.
- So there is some "redundancy" in $\mathcal{G}_2$.
- What about $\mathcal{G}_1$? Can we take out an edge and preserve the I-map?
- For example, omit the edge $X_2 \to X_3$ to get the graph $\mathcal{G}_1'$:

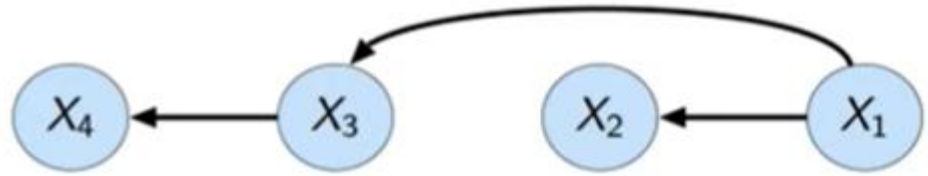
$$(X_3 \perp X_2 \mid X_1) \in \mathcal{I}_l(\mathcal{G}_1')$$



Minimal I-map

- Note that if the edges $X_2 \to X_4$ and $X_1 \to X_4$ are omitted from $\mathcal{G}_2$, we obtain $\mathcal{G}_1$, which is still an I-map for P.
- So there is some "redundancy" in $\mathcal{G}_2$.
- What about $\mathcal{G}_1$? Can we take out an edge and preserve the I-map?
- For example, omit the edge $X_2 \to X_3$ to get the graph $\mathcal{G}_1'$:

$$(X_3 \perp X_2 \mid X_1) \in \mathcal{I}_l(\mathcal{G}_1')$$



$$(X_3 \perp X_2 \mid X_1) \notin \mathcal{I}(P) \Longrightarrow \mathcal{G}_1' \text{ is } \textbf{not} \text{ an I-map for } P$$

Exercise: Show that if *any* edge of $\mathcal{G}_1$ is omitted, the result is not an I-map for P.

We say that $\mathcal{G}_1$ is a **minimal** I-map for P.

Minimal I-map

A graph $\mathcal{G}$ is a *minimal I-map* for P if it is an I-map for P, and if the removal of any edge from $\mathcal{G}$ makes it not an I-map.

D - Seperation

Case 1: Chain (A → B → C)

•**Example**: Rain → Wet Grass → Slippery Sidewalk.

•**Blocked if**: You know the middle node (B).

- *Why?* Once you know the grass is wet (B), rain (A) tells you nothing new about the sidewalk (C).

Case 2: Fork (A ← B → C)

•**Example**: Ice Cream Sales ← Hot Weather → Sunburns.

•**Blocked if**: You know the shared cause (B).

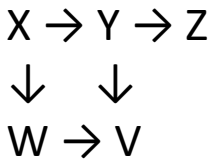
- *Why?* Knowing it's hot (B) explains both ice cream sales (A) and sunburns (C), so they're unrelated.

Case 3: Collider (A → B ← C)

•**Example**: Car Battery Dead → Car Won't Start ← Out of Gas.

•**Blocked if**: You *don't know* the shared effect (B) or its descendants.

- *Why?* Normally, battery (A) and gas (C) are unrelated. But if you know the car won't start (B), they become linked!



1.Are X and Z necessarily independent given evidence about Y?
Yes. The path from X to Z is X → Y → Z. Conditioning on Y blocks this path, making X and Z independent.

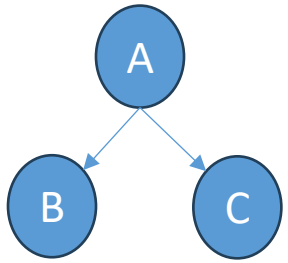
2.Are Y and V necessarily independent given evidence about W?
No. Y influences V through two paths:

1. Y → V (direct)
2. Y → W → V (indirect).
   Conditioning on W blocks the indirect path, but the direct path remains active. Thus, Y and V are still dependent.

3.Are X and V necessarily independent given no evidence?
No. X influences V through two paths:

1. X → Y → V
2. X → W → V.
   Without conditioning, both paths are active, so X and V are dependent.

A	B	C	Count	Log Likelihood
0	0	0	2	2 * ln (0.5 * 0.6 * 0.4) = -4.2406
0	0	1	1	1 * ln (0.5 * 0.6 * 0.6) = -1.7148
0	1	1	2	2 * ln (0.5 * 0.4 * 0.6) = -4.2406
1	0	0	1	1 * ln (0.5 * 0.2 * 0.6) = -2.8134
1	1	0	2	2 * ln (0.5 * 0.8 * 0.6) = -2.8542
1	1	1	2	2 * ln (0.5 * 0.8 * 0.4) = -3.6652



$P(A,B,C) = P(A).\ P(B|A)\ P(C|A)$:

$(A=0)=0.5$, $(A=1)=0.5$ **(1 parameter)**

$P(B|A)$: Estimated from counts **(2 parameters)**

- $(B=0|A=0)=0.6$
- $(B=1|A=0)=0.4$.
- $P(B=0|A=1)=0.2$
- $P(B=1|A=1)=0.8$.

$P(C|A)$: Estimated from counts **(2 parameters)**

- $P(C=0|A=0)=0.4$
- $P(C=1|A=0)=0.6$
- $P(C=0|A=1)=0.6$
- $P(C=1|A=1)=0.4$.

$$\text{BIC} = \ln(P(D|G)) - \frac{Parameters}{2}\ln(n)$$

n -> Number of records (10 in our example)

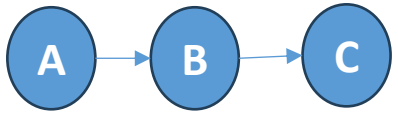
$Parameters$ -> Number of parameters in joint probability (5 in our example)

log$P(D|G)$ = -4.2406 -1.7148 -4.2406 -2.8134 -2.8542 -3.6652 ≈ -19.5288 ≈ -19.53

$$\text{Penalty} = \frac{Parameters}{2}ln(n) = \frac{5}{2}ln(10) = 5.75$$

BIC $= -19.53 - 5.75 =$ −25.28 s

We will select the model whose BIC is maximum



BIC= −24.62

Markov Model

MARKOV MODEL

Sequential Data:

- In many real-world applications, *data are sequential in nature*. This could be time-series data, spoken language, or even sensor readings from a robot.
- A key aspect of *sequential data* is that the *order of observations matters*. That is, the value of *a current observation often depends on previous ones*.
- Naive Approach: Ignore the Sequence
- One simplistic way to deal with sequential data is to *ignore its sequential nature* and assume that each observation is *independent* of the others as shown below.

MARKOV MODEL

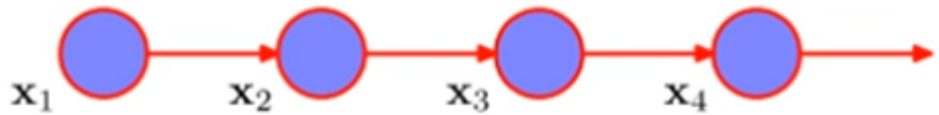
- What does independence mean?
- Suppose we have a *sequence of observations* ($x_1, x_2, \ldots, x_N$) which might be *temperature readings, stock prices, or sensor data* over time.
- If we treat them as *independent* observations, we assume that the *value of* x_i *does not depend on* x_{i-1} or any other previous value in the sequence.
- Mathematically, *the joint probability distribution of these observations* would simply be:

$$p(x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} p(x_i)$$

- This means that we *just multiply the probabilities of each individual observation*, assuming none of them affect each other.
- But in many cases, *assuming independence is unrealistic*. For example, the stock price today is often influenced by the stock price yesterday. Similarly, the temperature tomorrow depends on today's temperature
- By *ignoring these sequential dependencies*, we *fail to capture important patterns* or trends in the data

MARKOV MODEL

- Markov Models: Sequential Dependencies
- Markov models introduce a way to *account for the dependence* between observations while *still keeping the model relatively simple.*
- First-Order Markov Chain



- Markov Property:
- The *key assumption of a Markov Model* is that *an observation depends only on a fixed number of previous observations*, rather than the entire history.
- In a *first-order Markov model,* each observation depends only on the *immediately previous observation.*
- The *Markov property* for a first-order model states

$$p(x_i|x_1, x_2, ..., x_{i-1}) = p(x_i|x_{i-1})$$

- This means that the *current observation x_i depends only on x_{i-1}* and not on any earlier observations.

MARKOV MODEL

- Joint Probability in a First-Order Model:
- For a sequence of N observations (x_1,x_2,...,x_N) the *joint probability of this sequence* can be written *using the Markov property* as:

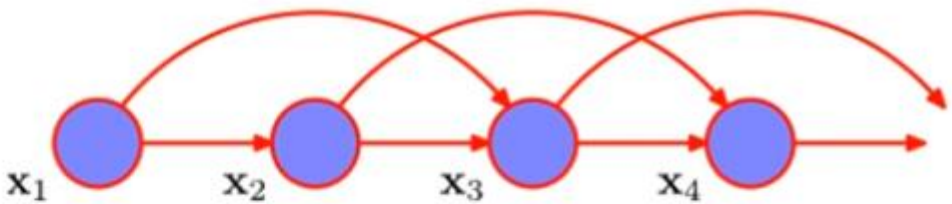
$$p(x_1, x_2, ..., x_N) = p(x_1) \prod_{i=2}^{N} p(x_i|x_{i-1})$$

- This formula means:
  - The probability of the *first observation $p(x_1)$ is independent* (since there's no earlier observation).
  - The probability of *each subsequent observation* x_i depends only on the observation immediately before it, x_{i-1}.
- This *factorization* makes the model computationally simpler compared to modeling dependencies on all previous observations.

MARKOV MODEL

- **Second-Order Markov Chain**

- In a *second-order Markov model*, each observation depends on the *two previous observations*. This adds more complexity but captures longer-range dependencies.



- The *Markov property for a second-order model* is

$$p(x_i|x_1, x_2, ..., x_{i-1}) = p(x_i|x_{i-1}, x_{i-2})$$

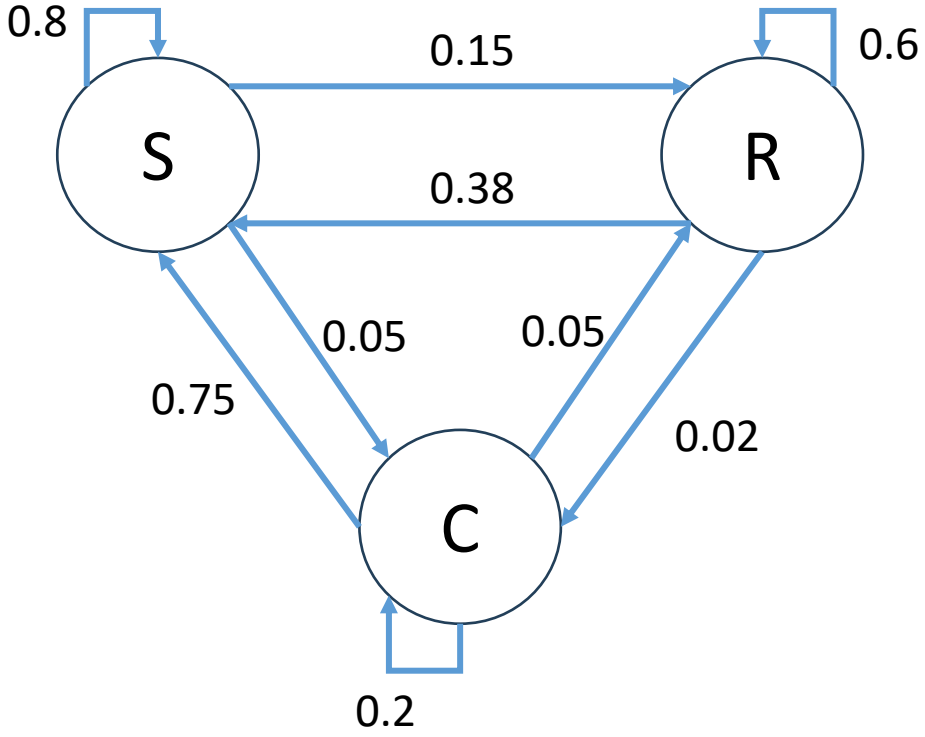
- This means that the current observation x_i depends on both x_{i-1} and x_{i-2}

MARKOV MODEL

- Joint Probability in a Second-Order Model:
- The *joint probability for a sequence of observations* under a second-order Markov model is

$$p(x_1, x_2, ..., x_N) = p(x_1)p(x_2|x_1)\prod_{i=3}^{N} p(x_i|x_{i-1}, x_{i-2})$$

- This means:
  - The probability of the first observation *$p(x_1)$ is independent.*
  - The probability of the second *observation $p(x_2|x_1)$ depends on the first observation*.
  - The probability of each subsequent *observation x_i depends on the two immediately preceding observations, x_{i-1} and x_{i-2}*



State Transition Diagram

State space or set of space

$$S = \{S, R, C\}$$

Initial State Distribution or Probability

$$\pi = \{0.7, 0.25, 0.05\}$$

$$P_{ij} = p(x_{t+1=j} \mid x_{t=i})$$

Transition Probability

	S	R	C
S	0.8	0.15	0.05
R	0.38	0.6	0.02
C	0.75	0.05	0.2

Q1) Given that the todays whether is Sunny what is the probability that tomorrow is sunny and day after tomorrow is Rainy?

$(t_1 = S, t_2 = S, t_3 = R)$

$P(t_3 = R, t_2 = S, t_1 = S) = P(t_3 \mid t_2) * P(t_2 \mid t_1)$ = 0.15 * 0.8 = 0.12

Q2) If today is cloudy and yesterday was Rainy. What is the probability that tomorrow would be sunny?

$(t_1 = S, t_2 = S, t_3 = R)$

$P(t_3 = R, t_2 = C, t_1 = S) = P(t_3 \mid t_2) * P(t_2 \mid t_1)$ = 0.75 * 0.02 =0.015

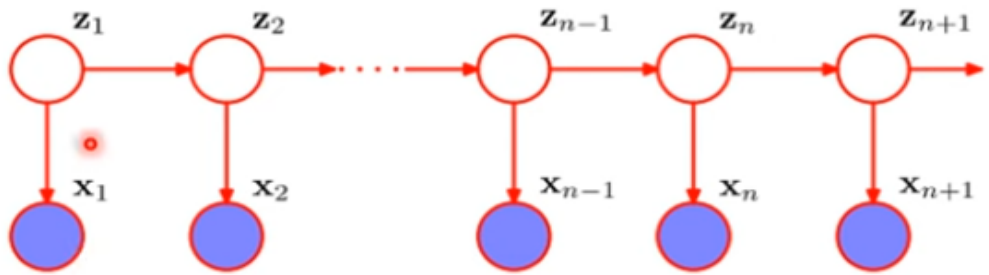
Probability of a Given Series:

S -> R -> R -> R -> C -> R

$P(S) * P(R \mid S) * P(R \mid R) * P(R \mid R) * P(C \mid R) * P(R \mid C)$

MARKOV MODEL

- Latent Variables:
- Another approach to simplifying the model is to introduce *latent (hidden) variables.*
- Instead of directly modeling the dependencies between the observations, we assume that there is *some hidden process* that generates the observations.
- This approach can *reduce the number of parameters* because the hidden variables often have simpler dependencies.



- The *blue circles* $x_1, x_2, \ldots, x_N$ represent the *observed data*. These could be real-world measurements like temperatures, stock prices, or sensor readings.
- The *red circles* represent *latent variables* $z_1, z_2, \ldots, z_N$, which are hidden variables that we don't observe directly but assume exist.

MARKOV MODEL

- The *latent variables* may represent some hidden state of the system that evolves over time and *influences the observed variables.*
- For *example*, in a weather system, the *latent variables could be the atmospheric conditions* that aren't directly measured but affect the observed temperatures.
- The *graphical structure with latent variables* represents a *Hidden Markov Model (HMM)*
- In this framework, the *joint probability of the observed sequence* $x_1, x_2, \ldots, x_N$ and *the latent sequence* $z_1, z_2, \ldots, z_N$ can be factored as:

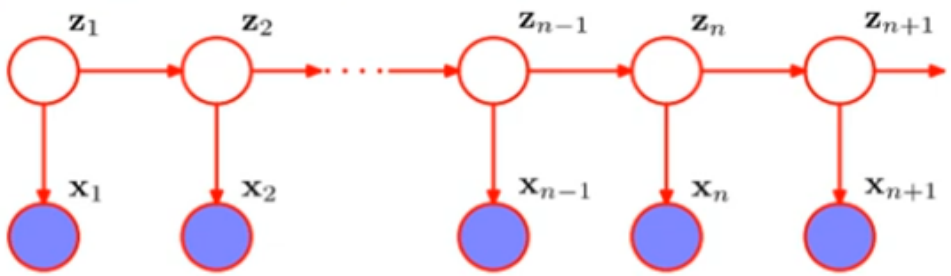
$$p(x_1, x_2, \ldots, x_N, z_1, z_2, \ldots, z_N) = p(z_1) \prod_{i=2}^{N} p(z_i|z_{i-1}) \prod_{i=1}^{N} p(x_i|z_i)$$

- This *expression breaks down* as:

1. $p(z_1)$: The probability of the first latent variable.
2. $\prod_{i=2}^{N} p(z_i|z_{i-1})$: The first-order Markov process governing the latent variables.
3. $\prod_{i=1}^{N} p(x_i|z_i)$: The conditional probability of each observed variable given its corresponding latent variable.

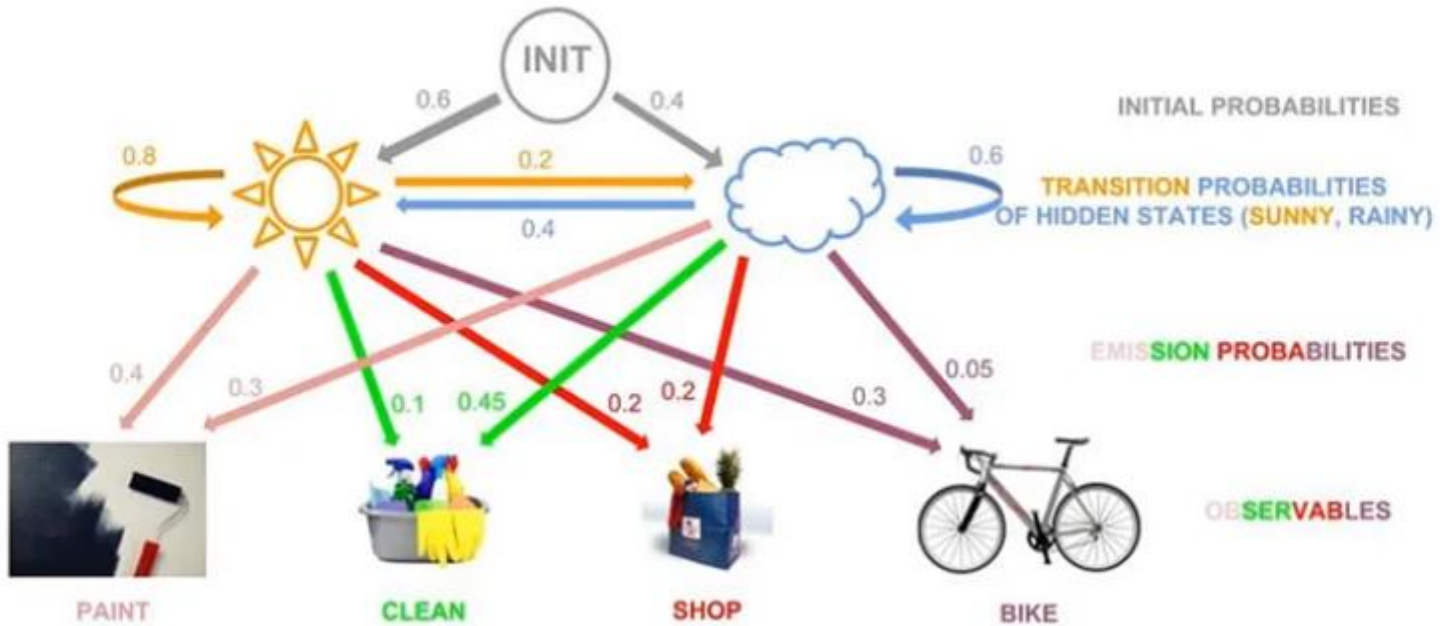
HIDDEN MARKOV MODELS

- *Hidden Markov Model (HMM)* is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (i.e. hidden) states.
- In probability theory, a *Markov model* is a stochastic model used to model randomly changing systems.
- It is assumed that *future states depend only on the current state*, not on the events that occurred before it (that is, it assumes the *Markov property*).

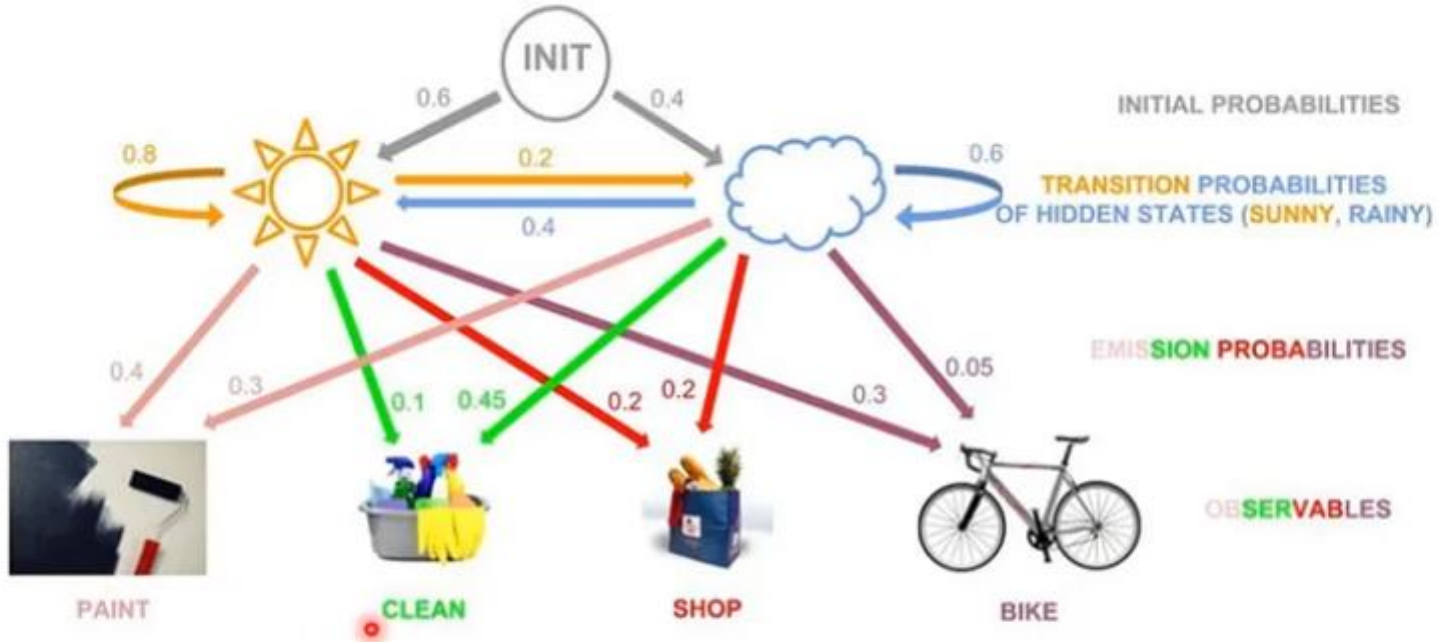


HIDDEN MARKOV MODELS

- Let's take Lisa our imaginary friend. *During the day she does either of these four things:*
- Painting
- Cleaning the house
- Biking
- Shopping for groceries
- From this *observation sequence* we want to know whether the day has been sunny or rainy. These two are going to be our *hidden states.*

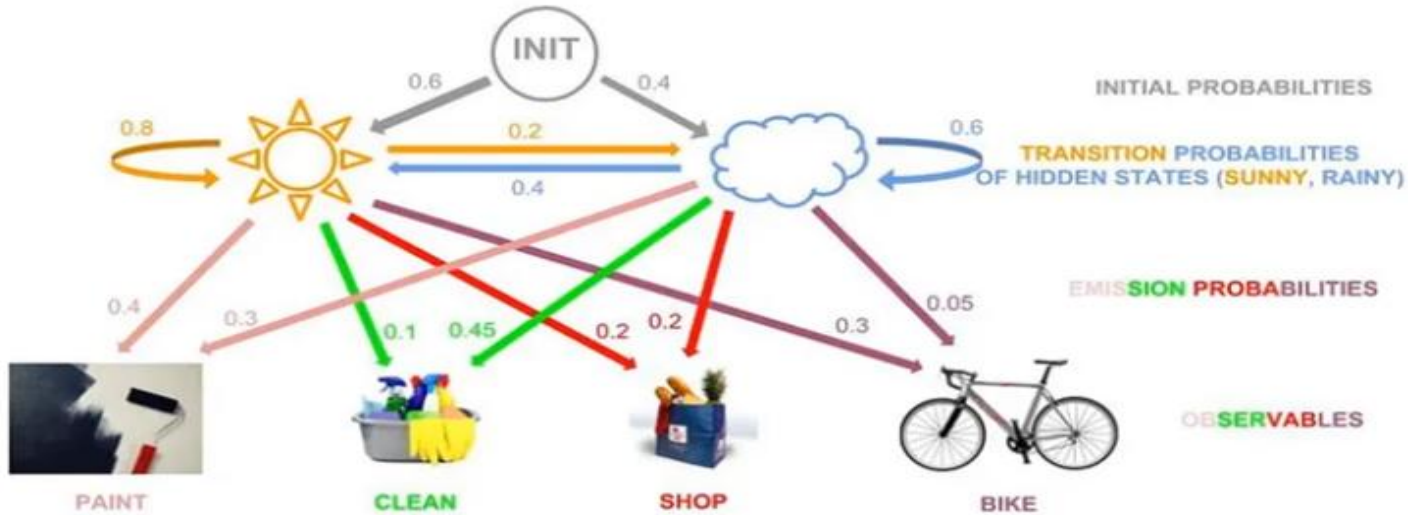


HIDDEN MARKOV MODELS



- To start off, a Hidden Markov Model consists of the following properties:
- Hidden States *S*: in the example above the hidden states are Sunny and Rainy, and they get grouped into a set S.
- Observables *O*: Paint, Clean, Shop and Bike. They get grouped into a set O.

HIDDEN MARKOV MODELS

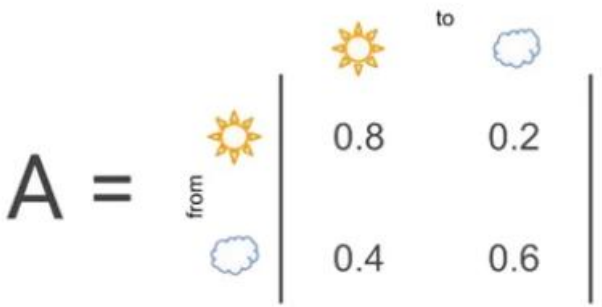


- **Initial Probabilities π:** a matrix of the initial likelihood of the state at time $t = 0$. In this case the likelihood that it is *Sunny on the first day is 0.6*, while the likelihood that it is *Rainy is 0.4*.
  $\pi = |0.6, 0.4|$

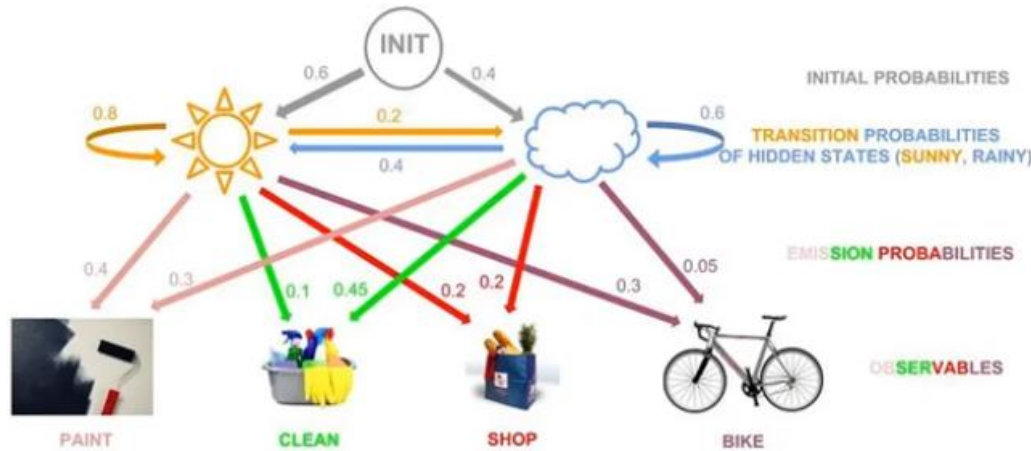
- *Note: every row of the following matrices must add up to 1 since they represent a probability.*

HIDDEN MARKOV MODELS

- Transition Probabilities *A:* a matrix that represents the probability of *transitioning to another state given the current state.*



$$A = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

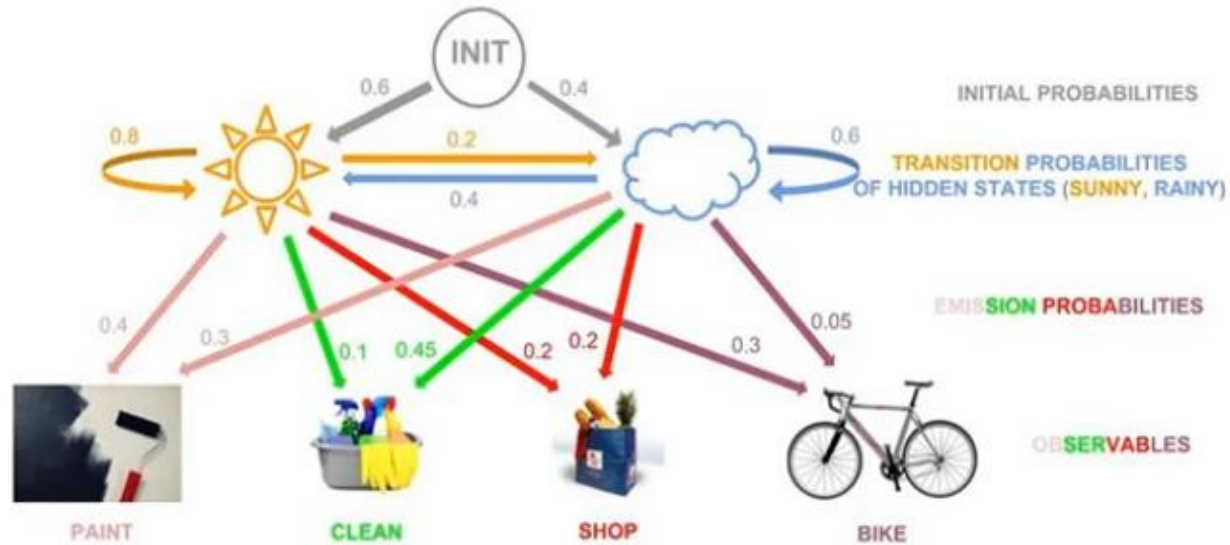


HIDDEN MARKOV MODELS

- Emission Probabilities *B:* a matrix that represents the probability of *seeing a specific observable given a hidden state*.

$$B = \begin{array}{c} \\ \text{from} \\ \\ \end{array} \begin{array}{c} \begin{array}{cccc} \text{PAINT} & \text{CLEAN} & \text{SHOP} & \text{BIKE} \end{array} \text{ to} \\ \left| \begin{array}{cccc} 0.4 & 0.1 & 0.2 & 0.3 \\ & & & \\ 0.3 & 0.45 & 0.2 & 0.05 \end{array} \right| \end{array}$$

$$B = \begin{vmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \end{vmatrix}$$



HIDDEN MARKOV MODELS

- Under *mathematical terms*, we would describe this model's properties as such:

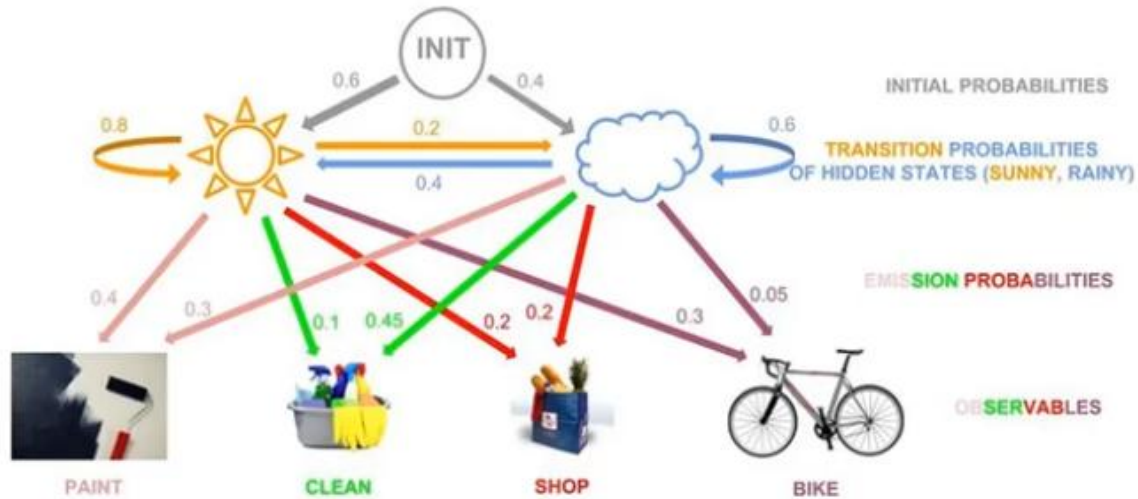
$$S = \{ S_{sunny}, S_{rainy} \} \quad (Hidden\ States)$$

$$O = \{ O_{clean}, O_{bike}, O_{shop}, O_{paint} \} \quad (Observables)$$

$$\pi = \begin{vmatrix} 0.6 & 0.4 \end{vmatrix} \quad (Initial\ Probabilities)$$

$$A = \begin{vmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{vmatrix} \quad (Transition\ Probabilities)$$

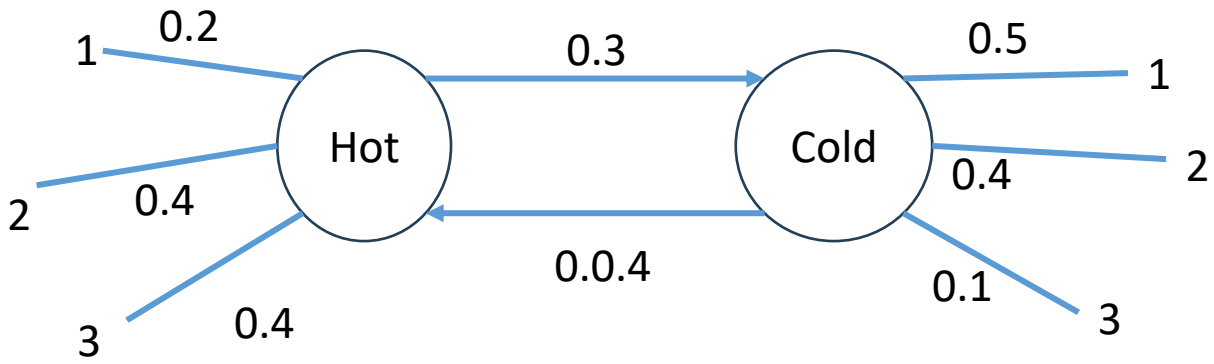
$$B = \begin{vmatrix} 0.4 & 0.1 & 0.2 & 0.3 \\ 0.3 & 0.45 & 0.2 & 0.05 \end{vmatrix} \quad (Emission\ Probabilities)$$



HIDDEN MARKOV MODELS

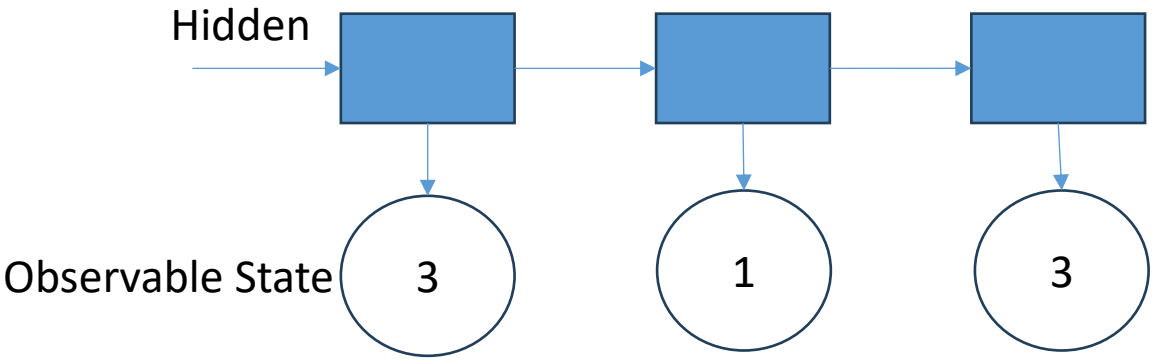
- **Challenges in HMM:**
- The Likelihood problem
- The Decoding problem

Decoding Viterbi Algorithm



Sum of outward arrow should be 1 so Hot to Hot is 0.7 and cold to cold will be 0.6

Observable State= $\{3,1,3\}$



N = No. Of Hidden State = 2

T = No. Of Observed State = 3

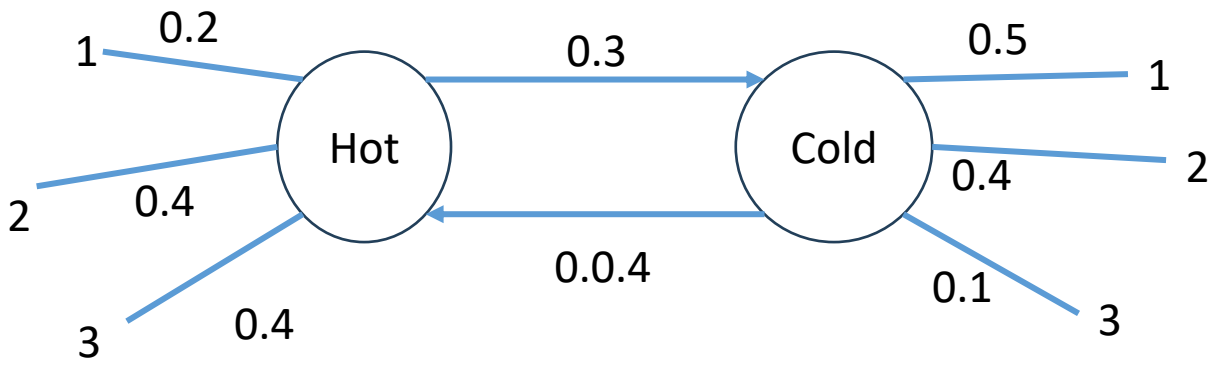
$M = N^T$

Transition Probability

	H	**C**
H	0.7	0.3
C	0.4	0.6

Emission Probability

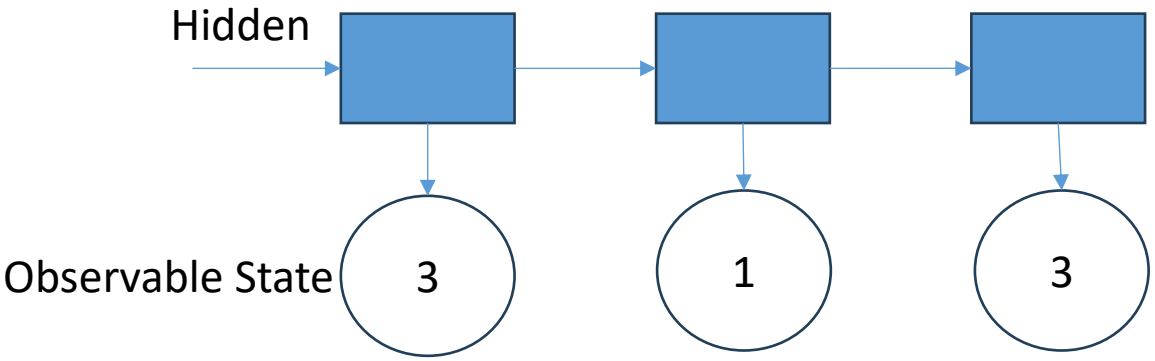
	1	**2**	**3**
H	0.2	0.4	0.4
C	0.5	0.4	0.1



Transition Probability

	H	C
H	0.7	0.3
C	0.4	0.6

Observable State= $\{3,1,3\}$



Emission Probability

	1	2	3
H	0.2	0.4	0.4
C	0.5	0.4	0.1

Initial State Distribution or Probability

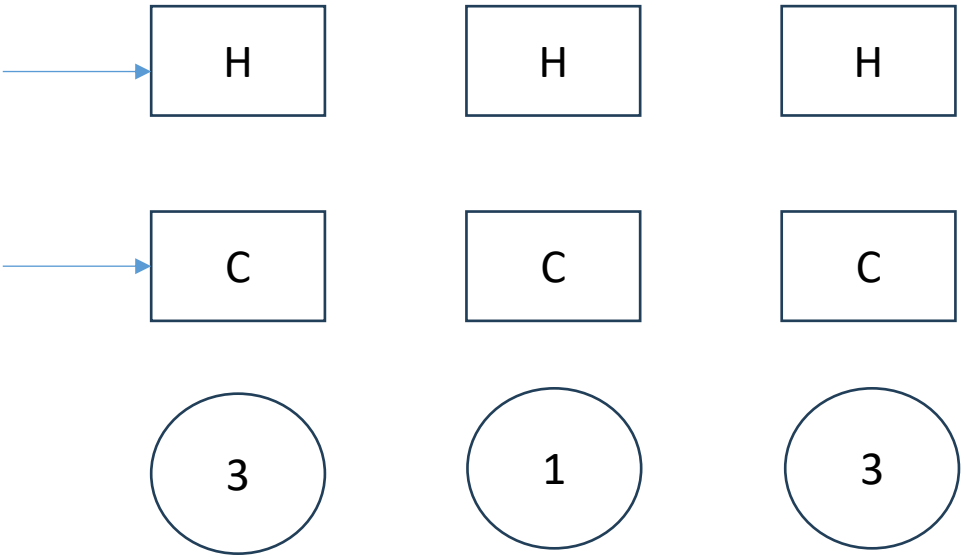
$$\pi = \{H = 0.8, \quad C = 0.2\}$$

$$M = N^T$$

N = Np. Of Hidden State = 2

T = Np. Of Observed State = 3

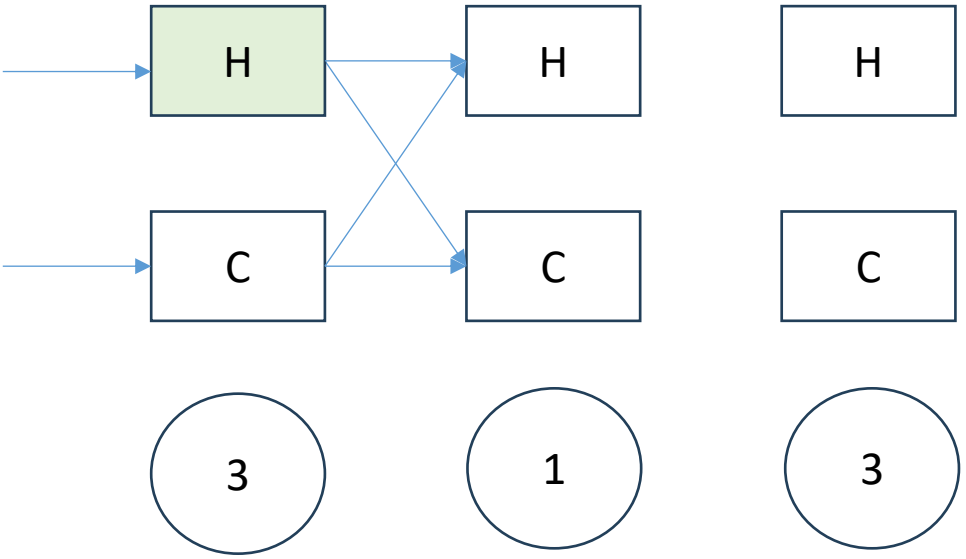
$$M = N^T$$



Step 1:

$V1(H) = P(3, H) = P(3 \mid H) * P(H)$ = 0.4 * 0.8 = 0.32

$V1(C) = P(3, C) = P(3 \mid C) * P(C)$ = 0.1 * 0.2 = 0.02



Step 2:

For hidden state H (Pick the Max)

$P(1,H) = P(1|H) * P(H|H) * P(H_{v1}) = 0.2 * 0.7 * 0.32 = 0.0448$ *# from Step 1 H*

$P(1,H) = P(1|H) * P(H|C) * P(C_{v1}) = 0.2 * 0.4 * 0.02 = 0.0016$ *# from Step 1 C*

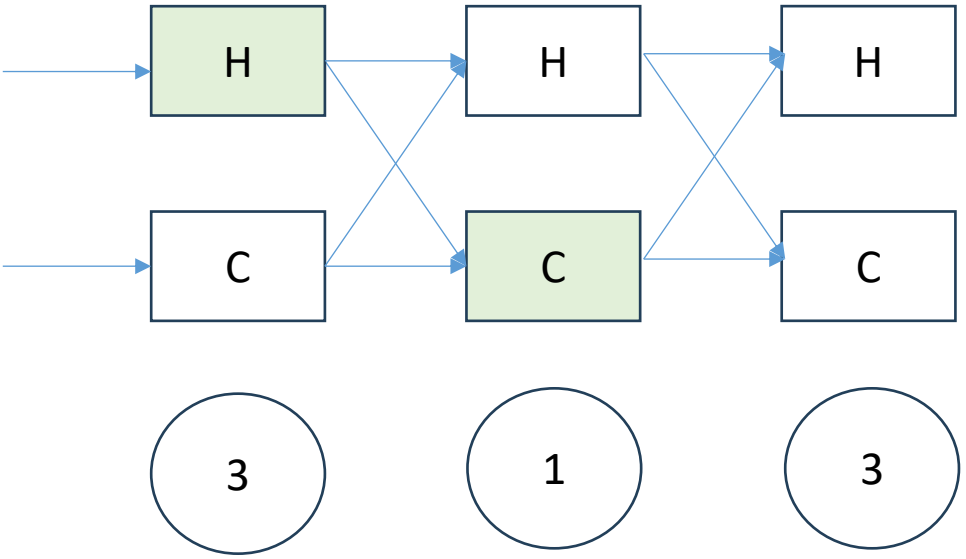
For hidden state C

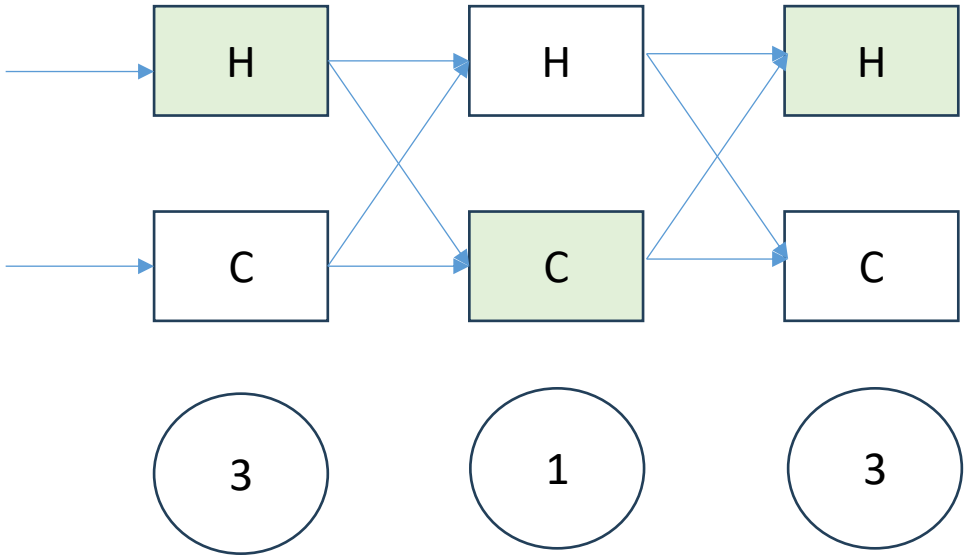
$P(1,C) = P(1|C) * P(C|H) * P(H_{v1}) = 0.5 * 0.3 * 0.32 = 0.048$ *# from Step 1 H*

$P(1,C) = P(1|C) * P(C|C) * P(C_{v1}) = 0.5 * 0.6 * 0.02 = 0.006$ *# from Step 1 C*

$V2(H) = 0.0448$

$V2(C) = 0.048$





Step 2:

For hidden state H (Pick the Max)

$P(3,H) = P(3|H) * P(H|H) * P(H_{v2}) = 0.4 * 0.7 * 0.0448 = 0.012544$ *# from Step 2 H*

$P(3,H) = P(3|H) * P(H|C) * P(C_{v2}) = 0.4 * 0.4 * 0.048 = 0.00768$ *# from Step 2 C*

For hidden state C

$P(3,C) = P(3|C) * P(C|H) * P(H_{S2}) = 0.1 * 0.3 * 0.0448 = 0.001344$ *# from Step 2 H*

$P(3,C) = P(3|C) * P(C|C) * P(C_{S2}) = 0.1 * 0.6 * 0.048 = 0.00288$ *# from Step 2 C*

$V3(H) = 0.012544$

$V3(C) = 0.00288$

- Likelihood Problem
- the *likelihood of a certain observation* sequence deriving from the HMM model
- Let's take the *initial example* about Lisa's activity in four days. The *observation sequence* is as follows: Paint, Clean, Shop and Bike.

$$O = \{ O_{paint}, O_{clean}, O_{shop}, O_{bike} \}$$

- So, what is the likelihood that this observation sequence O can derive from our HMM λ?

$$P(O|\lambda) = ???$$

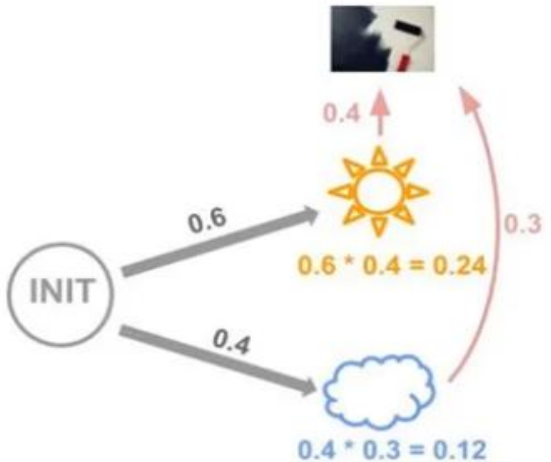
- There are *two methods* with which we can calculate this:
  - Forward Algorithm and the Backward Algorithm.

FORWARD ALGORITHM

- The Forward Algorithm
- The Forward algorithm comprises of three steps:
- Initialization
- Recursion
- Termination
- Initialization (at Time 1)

$$\alpha_1(i) = \pi_i \cdot b_i(O_1)$$

- The above equation means that the first forward variable is calculated by multiplying the initial probability of state i by the emission probability b of that state given the observable O at time 1.



$$B = \begin{array}{c|cccc} & \text{PAINT} & \text{CLEAN} & \text{SHOP} & \text{BIKE} \\ \hline \text{Sunny} & 0.4 & 0.1 & 0.2 & 0.3 \\ \text{Rainy} & 0.3 & 0.45 & 0.2 & 0.05 \end{array}$$

(from \ to)

$$B = \begin{vmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \end{vmatrix}$$

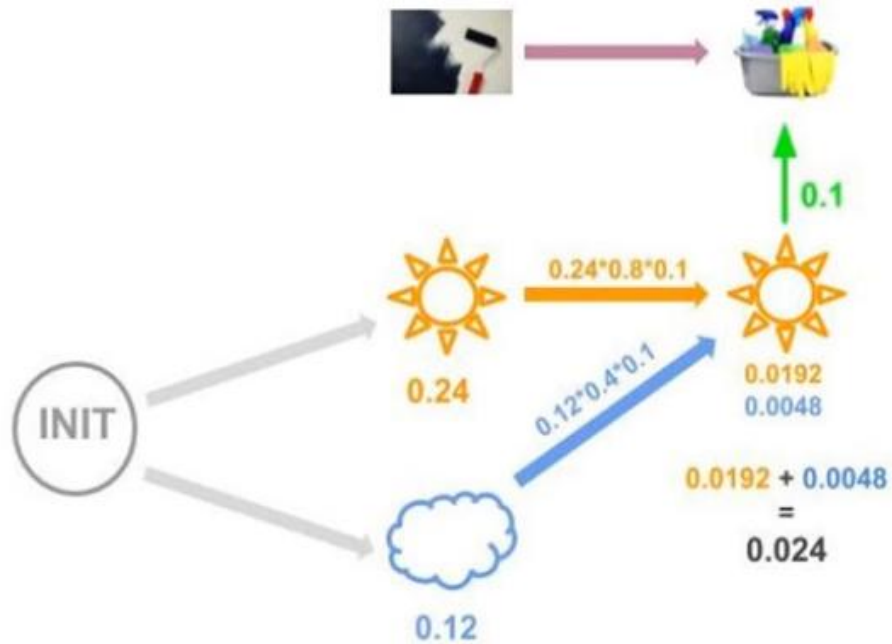
- Sunny:
- $\alpha_1(\text{Sunny}) = \boldsymbol{\pi}_1(\text{Sunny}) \times B(\text{Paint}|\text{Sunny})$
  $= 0.6 \times 0.4 = 0.24$
- Rainy:
- $\alpha_1(\text{Rainy}) = \pi_1(\text{Rainy}) \times B(\text{Paint}|\text{Rainy})$
  $= 0.4 \times 0.3 = 0.12$

FORWARD ALGORITHM

- Recursion (Sunny at Time 2)

$$\alpha_{t+1}(j) = \sum_{i=1}^{N} [\alpha_t(i) \cdot a_{ij}] \cdot b_j(O_{t+1})$$

- $\alpha_2(\text{Sunny}) = [\alpha_1(\text{Sunny}) \times A(\text{Sunny} \rightarrow \text{Sunny}) + \alpha_1(\text{Rainy}) \times A(\text{Rainy} \rightarrow \text{Sunny})] \times B(\text{Clean}|\text{Sunny})$
- $\alpha_2(\text{Sunny}) = [0.24 \times 0.8 + 0.12 \times 0.4] \times 0.1$
- $\alpha_2(\text{Sunny}) = [0.192 + 0.048] \times 0.1 = 0.024$



A =

from \ to		
	0.8	0.2
	0.4	0.6

$$A = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

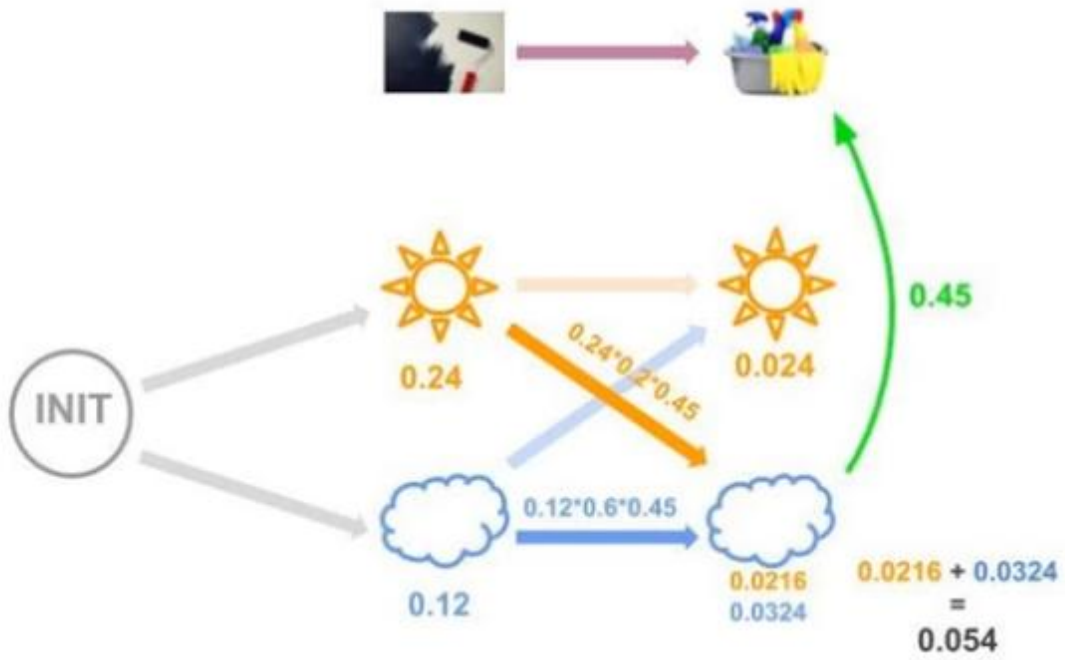
B =

from \ to	PAINT	CLEAN	SHOP	BIKE
	0.4	0.1	0.2	0.3
	0.3	0.45	0.2	0.05

$$B = \begin{vmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \end{vmatrix}$$

FORWARD ALGORITHM

- Similarly, for the next step we'll have a forward variable of 0.054 for the Rainy state:



$$\alpha 2(\text{Rainy}) = [\alpha_1(\text{Sunny}) \times A(\text{Sunny} \rightarrow \text{Rainy}) + \alpha_1(\text{Rainy}) \times A(\text{Rainy} \rightarrow \text{Rainy})] \times B(\text{Clean}|\text{Rainy})$$

$$= [0.24 * 0.2 + 0.12 * 0.6] * 0.45$$

$$= 0.12 \times 0.45 = 0.054$$

$$A = \begin{array}{c|cc} & \text{to} & \\ \text{from} & & \\ \hline & 0.8 & 0.2 \\ & 0.4 & 0.6 \end{array}$$

$$A = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

$$B = \begin{array}{c|cccc} & \text{to} & & & \\ \text{from} & \text{PAINT} & \text{CLEAN} & \text{SHOP} & \text{BIKE} \\ \hline & 0.4 & 0.1 & 0.2 & 0.3 \\ & 0.3 & 0.45 & 0.2 & 0.05 \end{array}$$

$$B = \begin{vmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \end{vmatrix}$$

FORWARD ALGORITHM

- Recursion at Time 3
- For the *third observable ("Shop")*, we again calculate the *forward probabilities* by considering transitions from time step 2.
- Sunny at Time 3:
- $\alpha_3(\text{Sunny})=[\alpha_2(\text{Sunny})\times A(\text{Sunny}\rightarrow\text{Sunny}) + \alpha_2(\text{Rainy})\times A(\text{Rainy}\rightarrow\text{Sunny})]\times B(\text{Shop}|\text{Sunny})$
- Substituting the values:
- $\alpha_3(\text{Sunny})=[0.024\times0.8+0.054\times0.4]\times0.2=0.01816$
- Rainy at Time 3:
- $\alpha_3(\text{Rainy})=[\alpha_2(\text{Sunny})\times A(\text{Sunny}\rightarrow\text{Rainy})+ \alpha_2(\text{Rainy})\times B(\text{Rainy}\rightarrow\text{Rainy})]\times B(\text{Shop}|\text{Rainy})$
- Substituting the values:
- $\alpha_3(\text{Rainy})=[0.024\times0.2+0.054\times0.6]\times0.2=0.00744$

$\alpha_2(\text{Sunny})= 0.024$
$\alpha_2(\text{Rainy}) = 0.054$

$A =$

from \ to		
	0.8	0.2
	0.4	0.6

$B =$

from \ to	PAINT	CLEAN	SHOP	BIKE
	0.4	0.1	0.2	0.3
	0.3	0.45	0.2	0.05

$$B = \begin{vmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \end{vmatrix}$$

FORWARD ALGORITHM

- Recursion at Time 4
- Finally, for the *fourth observable ("Bike")*, we calculate the forward probabilities for the last time step.
- Sunny at Time 4:
- $\alpha_4(\text{Sunny})=[\alpha_3(\text{Sunny})\times A(\text{Sunny}\rightarrow\text{Sunny})$
  $+\alpha_3(\text{Rainy})\times A(\text{Rainy}\rightarrow\text{Sunny})]\times B(\text{Bike}|\text{Sunny})$
- Substituting the values:
- $\alpha_4(\text{Sunny})=[0.00816\times0.8+0.00744\times0.4]\times0.3 = 0.00028512$
- Rainy at Time 4:
- $\alpha_4(\text{Rainy})=[\alpha_3(\text{Sunny})\times A(\text{Sunny}\rightarrow\text{Rainy})+$
  $\alpha_3(\text{Rainy})\times B(\text{Rainy}\rightarrow\text{Rainy})]\times B(\text{Bike}|\text{Rainy})$
- Substituting the values:
- $\alpha_4(\text{Rainy})=[0.00816\times0.2+0.00744\times0.6]\times0.05 = 0.0003048$

$\alpha_3(\text{Sunny}) = 0.00816$
$\alpha_3(\text{Rainy}) = 0.00744$

$A =$ (from \ to)

from \ to	☀	☁
☀	0.8	0.2
☁	0.4	0.6

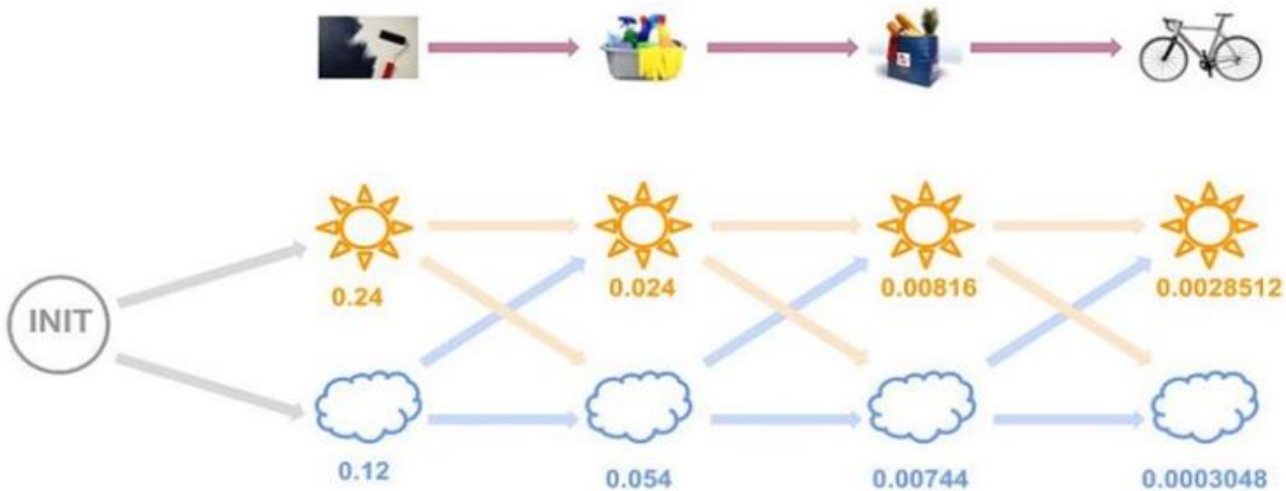
$B =$ (from \ to)

from \ to	PAINT	CLEAN	SHOP	BIKE
☀	0.4	0.1	0.2	0.3
☁	0.3	0.45	0.2	0.05

$$B = \begin{vmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \end{vmatrix}$$

FORWARD ALGORITHM

- We have all the forward variables:



- Termination

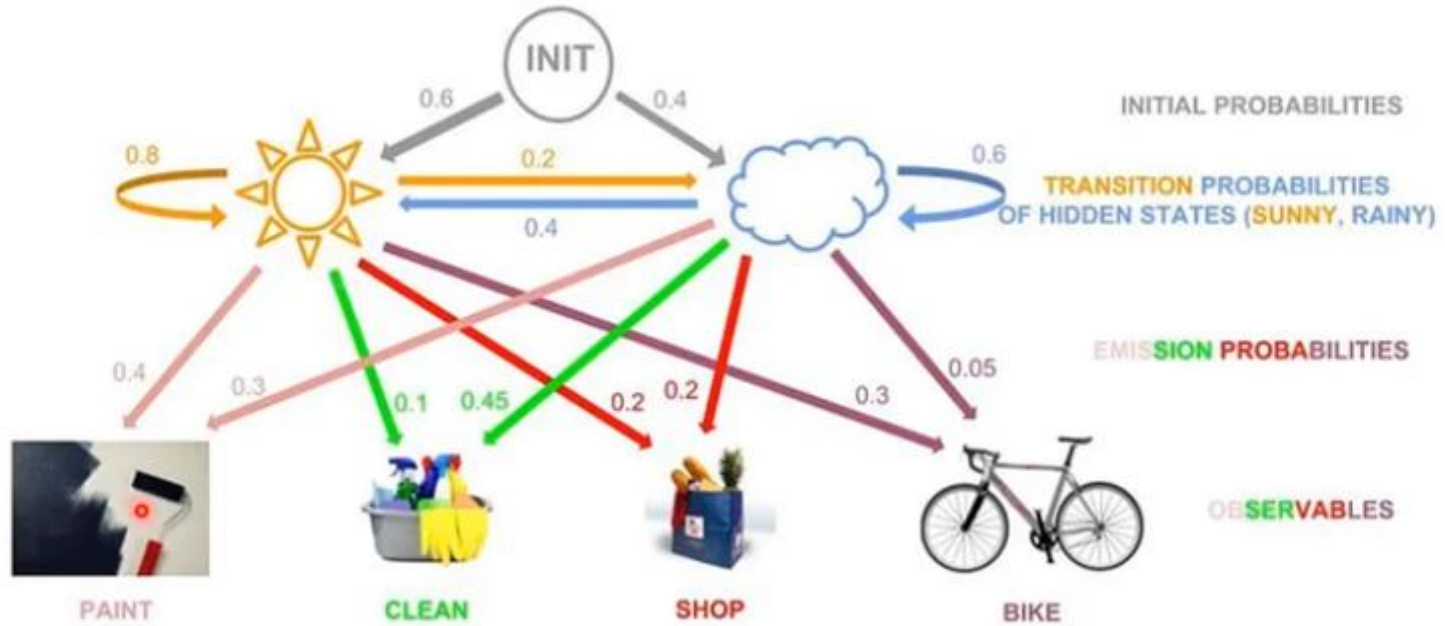
$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

- This final equation tells us that to find the probability of an observation sequence O deriving from an HMM model λ, we need to sum up all the forward variables at time T,
- i.e. all the variables of every state at the end of the observation sequence. Hence, in our example above,

$$P(O|\lambda) = 0.0028512 + 0.0003048 = 0.003156$$

HMM – BACKWARD ALGORITHM

- Let's take Lisa our imaginary friend. *During the day she does either of these four things:*
- Painting
- Cleaning the house
- Shopping for groceries
- Biking
- From this *observation sequence* we want to know whether the day has been sunny or rainy. These two are going to be our *hidden states.*

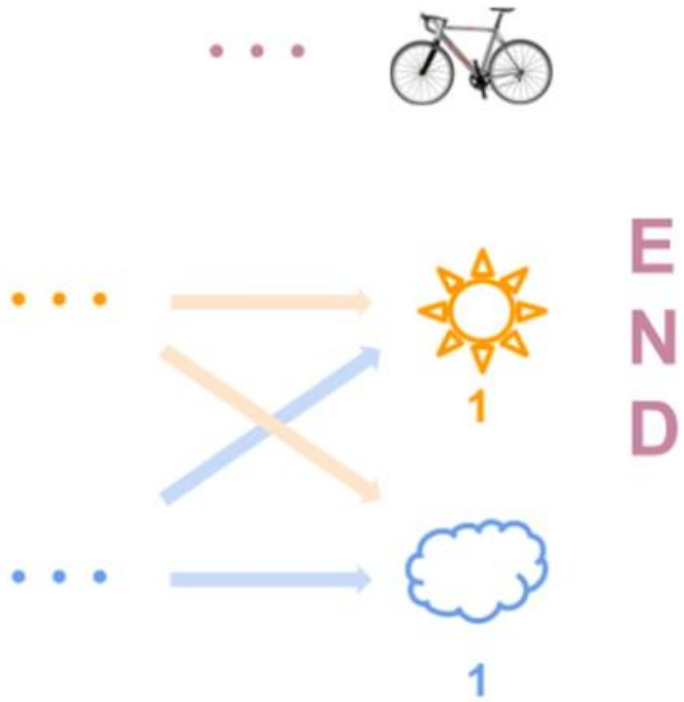


HMM – BACKWARD ALGORITHM

- Backward Algorithm
- The Backward Algorithm is similar to the Forward algorithm, but like the name *suggests it goes backward in time.*
- Steps involved are *Initialization, a Recursion and a Termination.*
- Initialization

$$\beta_T(i) = 1$$

- This equation is telling us is that at time T (at the end of the observation sequence) the *backward variables of every state is equal to 1.*

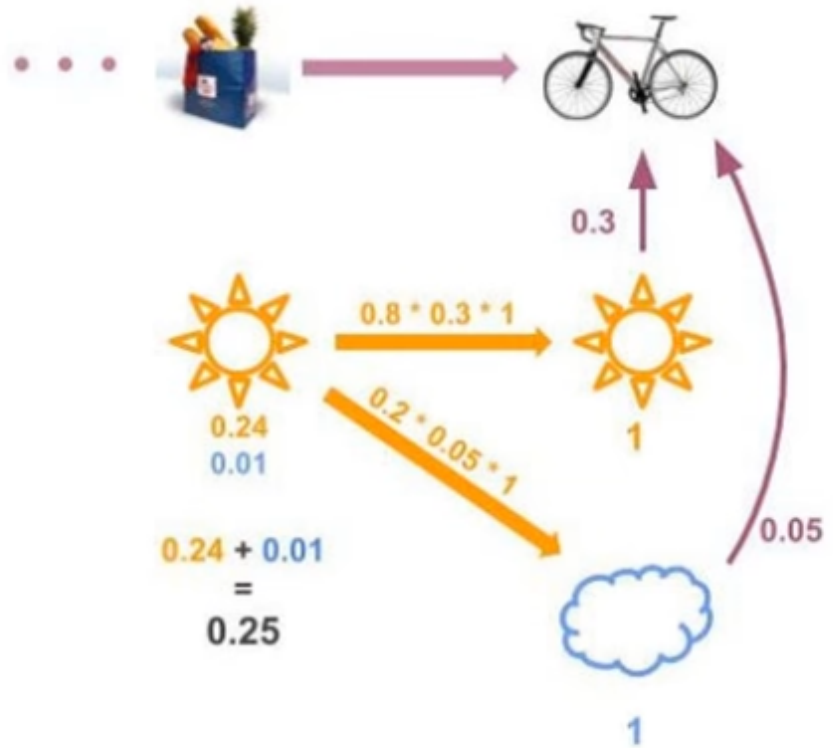


- β_4(Sunny)=1
- β_4(Rainy)=1

HMM – BACKWARD ALGORITHM

- **Recursion**

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} \cdot b_j(O_{t+1}) \cdot \beta_{t+1}(j)$$



- **For Sunny at t = 3:**
- $\beta_3(\text{Sunny}) = A(\text{Sunny} \rightarrow \text{Sunny}) \times B(\text{Bike}|\text{Sunny}) \times$

  $\beta_4(\text{Sunny}) + A(\text{Sunny} \rightarrow \text{Rainy}) \times B(\text{Bike}|\text{Rainy}) \times \beta_4(\text{Rainy})$

- **Substituting values:**
- $\beta_3(\text{Sunny}) = 0.8 \times 0.3 \times 1 + 0.2 \times 0.05 \times 1$
- $\beta_3 = 0.24 + 0.01 = 0.25$

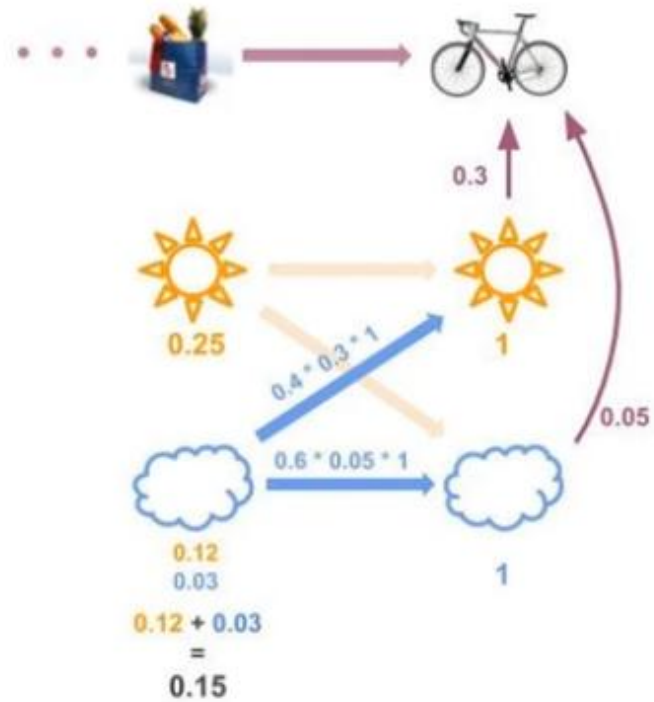
$$A = \begin{array}{c|cc} \text{from} \backslash \text{to} & \text{Sunny} & \text{Rainy} \\ \hline \text{Sunny} & 0.8 & 0.2 \\ \text{Rainy} & 0.4 & 0.6 \end{array} \qquad B = \begin{array}{c|cccc} \text{from} \backslash \text{to} & \text{PAINT} & \text{CLEAN} & \text{SHOP} & \text{BIKE} \\ \hline \text{Sunny} & 0.4 & 0.1 & 0.2 & 0.3 \\ \text{Rainy} & 0.3 & 0.45 & 0.2 & 0.05 \end{array}$$

$$A = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \qquad B = \begin{vmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \end{vmatrix}$$

HMM – BACKWARD ALGORITHM

- Recursion

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} \cdot b_j(O_{t+1}) \cdot \beta_{t+1}(j)$$



- For Rainy at t = 3:

- β_3(Rainy)=A(Rainy→Sunny)×B(Bike|Sunny)×

  β_4(Sunny)+A(Rainy→Rainy)×B(Bike|Rainy)× β_4(Rainy)

- Substituting values:

- β_3(Rainy)=0.4 × 0.3 × 1+ 0.6 × 0.05 × 1

- β_3 = 0.12 +0.03 = 0.15

A =

from \ to		
	0.8	0.2
	0.4	0.6

$$A = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

B =

from \ to	PAINT	CLEAN	SHOP	BIKE
	0.4	0.1	0.2	0.3
	0.3	0.45	0.2	0.05

$$B = \begin{vmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \end{vmatrix}$$

HMM – BACKWARD ALGORITHM

- Recursion
- For Sunny at t = 2:
- $\beta_2(\text{Sunny}) = A(\text{Sunny} \rightarrow \text{Sunny}) \times B(\text{Shop}|\text{Sunny}) \times \beta_3(\text{Sunny})$

  $+A(\text{Sunny} \rightarrow \text{Rainy}) \times B(\text{Shop}|\text{Rainy}) \times \beta_3(\text{Rainy})$

  $= 0.8 * 0.2 * 0.25 + 0.2 * 0.2 * 0.15$

  $= 0.04 + 0.006 = \mathbf{0.046}$

- For Rainy at t = 2:
- $\beta_2(\text{Rainy}) = A(\text{Rainy} \rightarrow \text{Sunny}) \times B(\text{Shop}|\text{Sunny}) \times \beta_3(\text{Sunny})$

  $+A(\text{Rainy} \rightarrow \text{Rainy}) \times B(\text{Shop}|\text{Rainy}) \times \beta_3(\text{Rainy})$

  $= 0.4 \times 0.2 \times 0.25 + 0.6 \times 0.2 \times 0.15$

  $= 0.02 + 0.018 = \mathbf{0.038}$

- $\beta_3(\text{Sunny}) = 0.25$
- $\beta_3(\text{Rainy}) = 0.15$

$A =$

from \ to	☀	☁
☀	0.8	0.2
☁	0.4	0.6

$B =$

from \ to	PAINT	CLEAN	SHOP	BIKE
☀	0.4	0.1	0.2	0.3
☁	0.3	0.45	0.2	0.05

$$B = \begin{vmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \end{vmatrix}$$

HMM – BACKWARD ALGORITHM

- Recursion
- For Sunny at t = 1:
- $\beta_1(\text{Sunny}) = A(\text{Sunny} \rightarrow \text{Sunny}) \times B(\text{Clean}|\text{Sunny}) \times \beta_2(\text{Sunny})$

  $+A(\text{Sunny} \rightarrow \text{Rainy}) \times B(\text{Clean}|\text{Rainy}) \times \beta_2(\text{Rainy})$

  $= 0.8 \times 0.1 \times 0.046 + 0.2 \times 0.45 \times 0.038$

  $= 0.00368 + 0.00342 = 0.0071$
- For Rainy at t = 1:
- $\beta_1(\text{Rainy}) = A(\text{Rainy} \rightarrow \text{Sunny}) \times B(\text{Clean}|\text{Sunny}) \times \beta_2(\text{Sunny})$

  $+A(\text{Rainy} \rightarrow \text{Rainy}) \times B(\text{Clean}|\text{Rainy}) \times \beta_2(\text{Rainy})$

  $= 0.4 \times 0.1 \times 0.046 + 0.6 \times 0.45 \times 0.038$

  $= 0.00184 + 0.01026 = 0.0121$

- $\beta_2(\text{Sunny}) = 0.046$
- $\beta_2(\text{Rainy}) = 0.038$

$A =$

from \ to	☀	☁
☀	0.8	0.2
☁	0.4	0.6

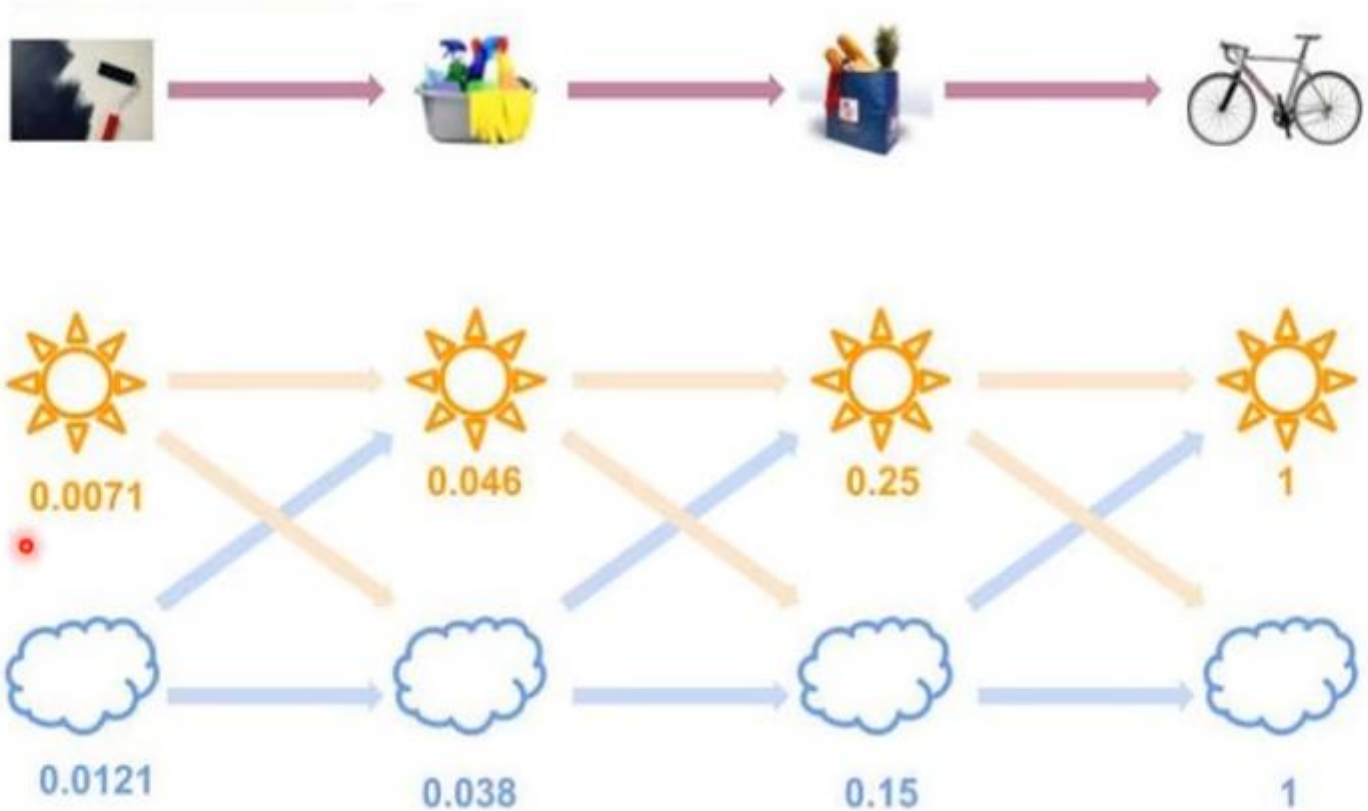
$B =$

from \ to	PAINT	CLEAN	SHOP	BIKE
☀	0.4	0.1	0.2	0.3
☁	0.3	0.45	0.2	0.05

$$B = \begin{vmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \end{vmatrix}$$

HMM – BACKWARD ALGORITHM

- Final Backward Probabilities



HMM – BACKWARD ALGORITHM

- Termination

$$P(O|\lambda) = \sum_{i=1}^{N} \pi_i \cdot b_i(O_1) \cdot \beta_1(i)$$

- Where:
- π_i is the *initial probability* of being in state (Sunny or Rainy).
- $b_i(O_1)$ is the *emission probability* of the first observation O_1
- β_1 is the *backward probability* at time t=1
- Initial probabilities:
- π(Sunny)=0.6 , π(Rainy)=0.4
- Emission probabilities for "Paint":
- B(Paint|Sunny)=0.4, B(Paint | Rainy) = 0.3

- Backward probabilities at t=1
- β1(Sunny)=0.0071
- β1(Rainy)=0.0121
- Substituting the values we get:
- Considering Sunny State:
- P(O) = π(Sunny)×B(Paint|Sunny)×β1(Sunny)

  = 0.6 * 0.4 * 0.0071 = 0.001704
- Considering Rainy State:
- P(O) = π(Rainy)×B(Paint|Rainy)×β1(Rainy)

  = 0.4 * 0.3 * 0.0121 = 0.001452
- Final value is:
- P(O|λ) = 0.001704 + 0.001452 = 0.003156