Advanced Artificial Intelligence

Angshuman Paul

Assistant Professor

Department of Computer Science & Engineering

॥ त्वं ज्ञानमयो विज्ञानमयोऽसि ॥ ॥ विवेकख्यातिरविप्लवा हानोपायः ॥

VISION INNOVATION EXCELLENCE

Uncertainty

- So far, we have learnt about deterministic scenarios
  - I have a specific clause that needs to satisfy
  - No notion of randomness
- In logic, we have to define a variable for each such possibilities
  - In many problems, there are a huge number of such possibilities
  - Example: a person has cough
    - What is the possible cause?
    - There may be a huge number of causes for this
    - It's almost impossible to enumerate all such possibilities
- We often have to make decision based on the possibilities

Sample Space

- The set of possible outcomes
- Rolling a dice
  - Sample space $S = \{1, 2, 3, 4, 5, 6\}$
  - Each element of sample space: sample point

- Sample space can be
  - Finite (above example)
  - Infinite (Amount of rainfall at Jodhpur in July)

Event

- Any subset of the sample space
- Example: In the context of rolling a dice
  - The outcome is a number < 3
  - Event $A = \{1, 2\}$

Random Variables

- In logic, we have symbols

- In probability, we have random variables (rv)
  - A numerical description of the outcomes of a random experiment
  - A function that assigns numerical values (real or Boolean) to each sample point
  - Usually indicated using capital letters (e.g., X)
    - Discrete: takes only a countable number of discrete values
      - Sample space for weather condition: $\{sunny, rainy, cloudy\}$
    - Continuous: takes uncountably infinite number of possible values
      - Sample space for temperature at Jodhpur: $[6.7^o, 46.2^o]$

Probability of an Event

- Consider an event A
- Probability of event A
  - $P(A) = \frac{Number\ of\ elements\ in\ set\ A}{Number\ of\ elements\ in\ the\ sample\ space\ S}$
  - $P(A) = \frac{Number\ of\ favourable\ outcomes}{Total\ number\ of\ outcomes}$
- Example: In the context of rolling a dice
  - Event A: The outcome is an odd number
  - Event $A = \{\ldots\}$

Probability of an Event

- Probability of event A
  - $P(A) = \frac{Number\ of\ elements\ in\ set\ A}{Number\ of\ elements\ in\ the\ sample\ space\ S}$
  - $P(A) = \frac{Number\ of\ favourable\ outcomes}{Total\ number\ of\ outcomes}$
- Example: In the context of rolling a dice
  - Event A: The outcome is an odd number
  - Event $A = \{1, 3, 5\}$

Probability of an Event

- Probability of event A
  - $P(A) = \frac{Number\ of\ elements\ in\ set\ A}{Number\ of\ elements\ in\ the\ sample\ space\ S}$
  - $P(A) = \frac{Number\ of\ favourable\ outcomes}{Total\ number\ of\ outcomes}$
- Example: In the context of rolling a dice
  - Event A: The outcome is an odd number
  - Event $A = \{1, 3, 5\}$
  - $P(A) = \frac{3}{6} = 0.5$

Probability of an Event

- Probability of event A
  - $P(A) = \frac{Number\ of\ elements\ in\ set\ A}{Number\ of\ elements\ in\ the\ sample\ space\ S}$
  - $P(A) = \frac{Number\ of\ favourable\ outcomes}{Total\ number\ of\ outcomes}$
- Example: In the context of rolling a dice
  - Event A: The outcome is an odd number
  - Event $A = \{1, 3, 5\}$
  - $P(A) = \frac{3}{6} = 0.5$

What is the problem with this definition?

Probability of an Event

- In an experiment with **finite sample space and equally likely outcomes**, Probability of event A
  - $P(A) = \frac{Number\ of\ elements\ in\ set\ A}{Number\ of\ elements\ in\ the\ sample\ space\ S}$
  - $P(A) = \frac{Number\ of\ favourable\ outcomes}{Total\ number\ of\ outcomes}$
- Example: In the context of rolling a dice
  - Event A: The outcome is an odd number
  - Event $A = \{1, 3, 5\}$
  - $P(A) = \frac{3}{6} = 0.5$

Probability of an Event

- In an experiment with **finite sample space and equally likely outcomes**, Probability of event A
  - $P(A) = \frac{Number\ of\ elements\ in\ set\ A}{Number\ of\ elements\ in\ the\ sample\ space\ S}$
  - $P(A) = \frac{Number\ of\ favourable\ outcomes}{Total\ number\ of\ outcomes}$
- Example: In the context of rolling a dice
  - Event A: The outcome is an odd number
  - Event $A = \{1, 3, 5\}$
  - $P(A) = \frac{3}{6} = 0.5$

What if this condition does not hold?

Frequentist Approach of Probability

- Suppose we do an experiment n number of times
- Out of these, event A occurs $n(A)$ number of times

- Relative frequency of A is $f_r(A) = \frac{n(A)}{n}$
- Probability of A is $P(A) =$

Frequentist Approach of Probability

- Suppose we do an experiment n number of times
- Out of these, event A occurs $n(A)$ number of times

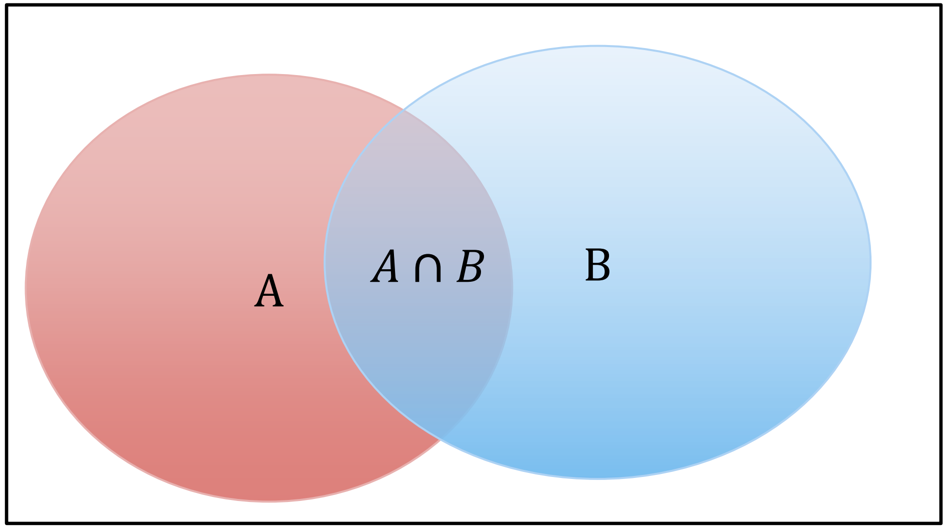
- Relative frequency of A is $f_r(A) = \frac{n(A)}{n}$
- Probability of A is $P(A) = \lim_{n \to \infty} f_r(A) = \lim_{n \to \infty} \frac{n(A)}{n}$

Probability

- A probability measure or probability function $P(\cdot)$ assigns a probability to an event
  - $P(A)$ is the chance that event A occurs

Conditional Probability

$$\boldsymbol{P(A|B) = \frac{P(A \cap B)}{P(B)}}$$

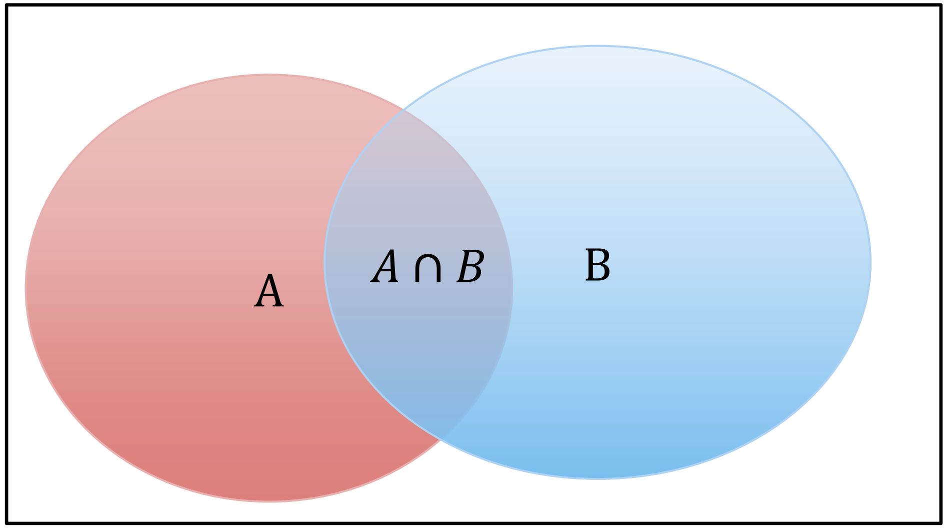


Conditional Probability

Prior Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)\,P(A)}{P(B)}$$

Indicates my belief about the occurrence of A

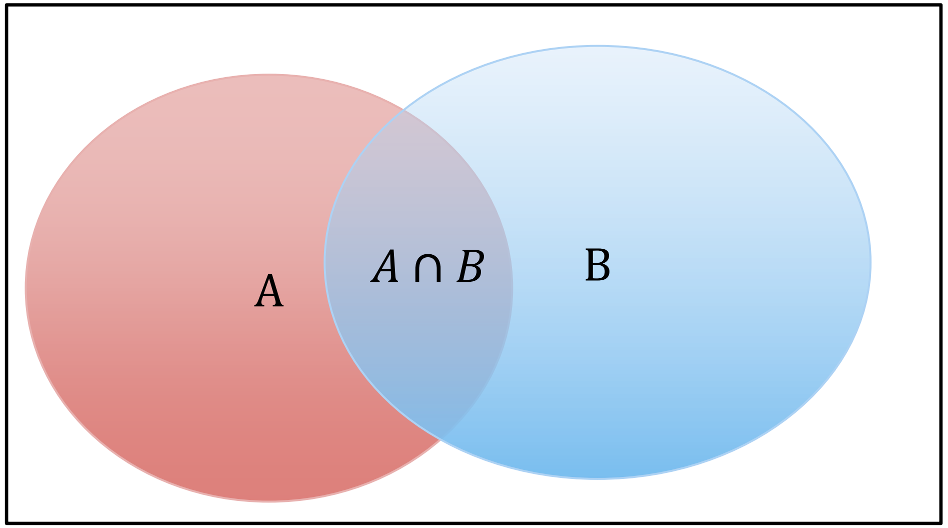


Conditional Probability

$$\boldsymbol{P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)\,P(A)}{P(B)}}$$

Likelihood

Indicates the chance of B to occur given that A has occurred

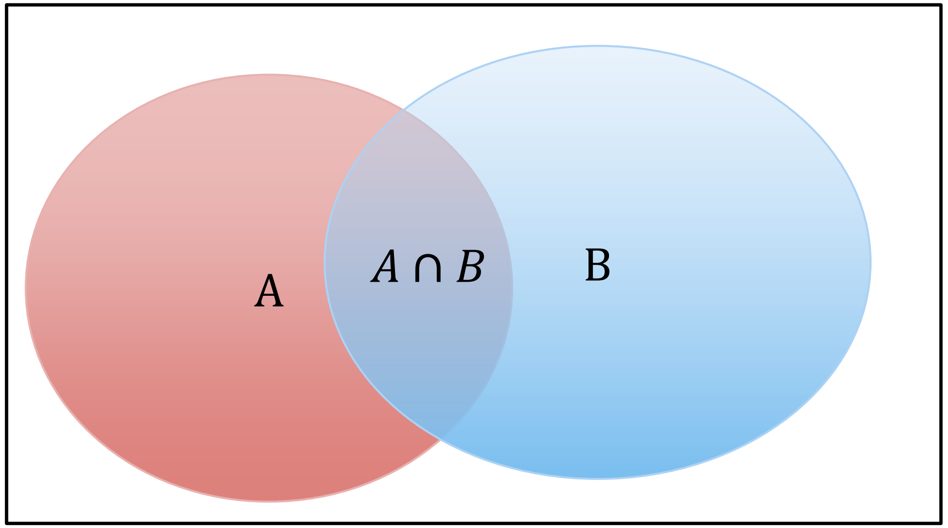


Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)\,P(A)}{P(B)}$$

Evidence

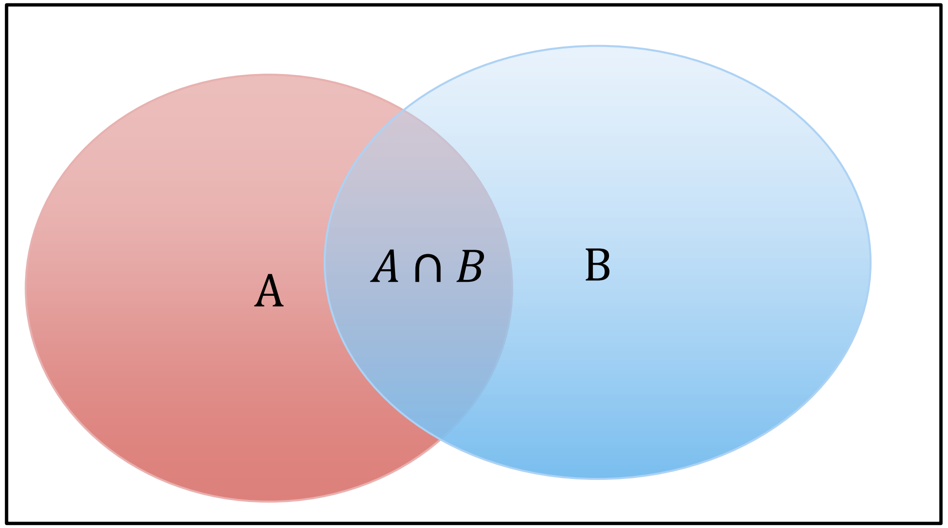
Probability that B occurs



Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)\,P(A)}{P(B)}$$

Posterior/
Conditional
probability

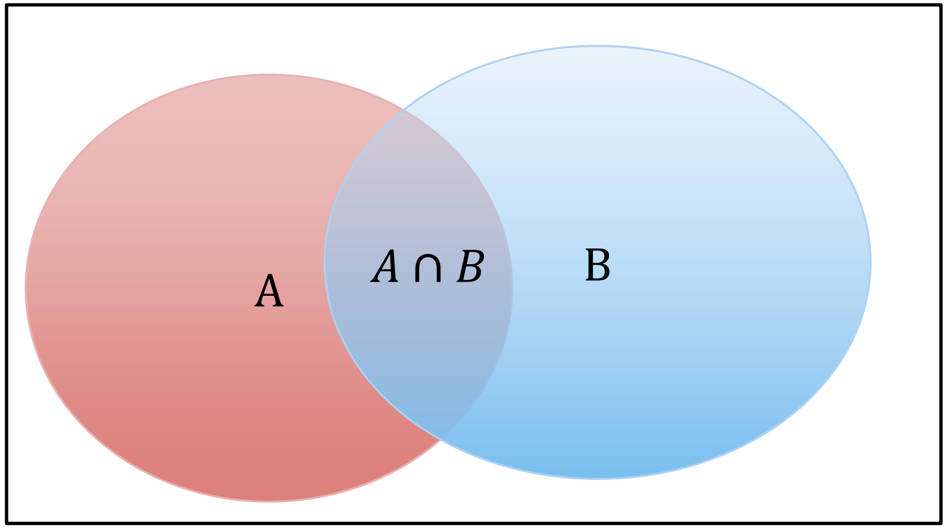


Conditioned on the evidence that we have seen

Conditional Probability

$$\boldsymbol{P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)\,P(A)}{P(B)}}$$

Bayes' Theorem



Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)\,P(A)}{P(B)}$$

Bayes' Theorem

$$P(Cause|Effect) = \frac{P(Effect|Cause)\,P(Cause)}{P(Effect)}$$

Conditional Probability

- Consider a system with causes c1, and c2, and effects e1, e2 and e3

- Suppose we want to find the probabilities of different causes given effect e2

$$\boldsymbol{P(Cause|Effect) = \frac{P(Effect|Cause)\,P(Cause)}{P(Effect)}}$$

Independent Events

- Two events A and B are said to be independent if
  - $\boldsymbol{P(A|B) = P(A)}$ (1)

- We already have
  - $\boldsymbol{P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)\,P(A)}{P(B)}}$ (2)

- From (1) and (2), we get
  - $\boldsymbol{P(B|A) = P(B)}$
  - $\boldsymbol{P(A \cap B) = P(A)P(B)}$

Joint Probability Distribution

- Consider two random variables
- X corresponding to weather $\{sunny, rainy, cloudy\}$
  - $P(X) = \{0.6, 0.1, 0.3\}$
- Y corresponding to power cut $\{power\ cut, no\ power\ cut\}$
  - $P(Y) = \{0.15, 0.85\}$
- A joint probability distribution of $X\ and\ Y$
  - Probability distribution on all possible pairs of outputs

Joint Probability Distribution

- X corresponding to weather $\{sunny, rainy, cloudy\}$
  - $P(X) = \{0.6, 0.1, 0.3\}$
- Y corresponding to power cut $\{power\ cut, no\ power\ cut\}$
  - $P(Y) = \{0.15, 0.85\}$
- A joint probability distribution of $X\ and\ Y$
  - Probability distribution on all possible pairs of outputs
- A 3×2 matrix of values

Chain Rule

- If $A_1, A_2, \dots, A_n$ are n events, then
  - $P(A_n \cap A_{n-1} \cap \cdots \cap A_1) = P(A_n | A_{n-1} \cap \cdots \cap A_1)\, P(A_{n-1} \cap \cdots \cap A_1)$ (1)
- Similarly,
  - $P(A_{n-1} \cap A_{n-2} \cap \cdots \cap A_1) = P(A_{n-1} | A_{n-2} \cap \cdots \cap A_1)\, P(A_{n-2} \cap \cdots \cap A_1)$ (2)
- Extending this for the subsequent events and putting in (1), we get,
  - $P(A_n \cap A_{n-1} \cap \cdots \cap A_1)$

    $= P\,(A_n | A_{n-1} \cap \cdots \cap A_1)\, P(A_{n-1} | A_{n-2} \cap \cdots \cap A_1) P(A_{n-2} | A_{n-3} \cap \cdots \cap A_1) \dots P(A_1)$

Chain Rule

- If A_1, A_2, A_3 are 3 events, then we use
  - $P(A_n \cap A_{n-1} \cap \cdots \cap A_1)$

    $= P\ (A_n | A_{n-1} \cap \cdots \cap A_1)\ P(A_{n-1} | A_{n-2} \cap \cdots \cap A_1) P(A_{n-2} | A_{n-3} \cap \cdots \cap A_1) \ldots P(A_1)$

- We get
  - $P(A_4 \cap A_3 \cap A_2 \cap A_1)$

    $= P\ (A_4 | A_3 \cap A_2 \cap A_1)\ P(A_3 | A_2 \cap A_1) P(A_2 | A_1) P(A_1)$

Markov Decision Process

- Many decision making problems happen with uncertainty but for a long time
- Decision theory: episodic (only one decision)
- In MDP, we focus on a long sequence of actions
- Stochastic transitions (Actions have stochastic outcome)
- Many probabilistic problems can be converted into MDP

Markov Decision Process

- A set of states $\mathcal{S}$
- A set of actions $\mathcal{A}$
- A transition function $\mathcal{T}(s, a, s')$: An agent in state s reaches state s' with probability $\mathcal{T}(s, a, s')$ if the agent performs action a
- A cost model $\mathcal{C}(s, a, s')$: If the agent reaches state s' from s by taking action a, $\mathcal{C}(s, a, s')$ is the cost that the agent pays

Markov Process

- Next state depends on the current state and not on the past (first order Markov process)

- $\mathcal{C}(s, a, s'), \mathcal{T}(s, a, s')$: first order Markov process
    - How we arrived at s does not matter to determine cost or the next transition

Markov Decision Process

- A set of goals $\mathcal{G}$ may be given: absorbing / non-absorbing goals
- A start state may be given s_0
- Discount factor γ may be given
- A reward model $\mathcal{R}(s, a, s')$ may be given

What is the Objective of Markov Decision Process?

- In search problems (deterministic), I find a path from a start state to goal state
- In MDP, I am looking for a table that tells me which action to take when the agent is at a particular state
  - Policy
- After the action is taken, the agent knows exactly where (next) it reaches
  - i.e. we assume full observability

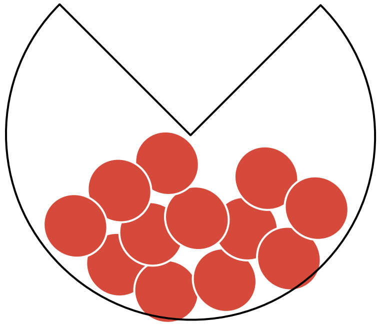
Markov Decision Process

- In search problems, we optimize cost to reach a goals
  - Since it is deterministic

- In MDP, we are going to optimize expected cost / expected rewards to reach the goal
  - Since my transitions are stochastic

Markov Model with an Example

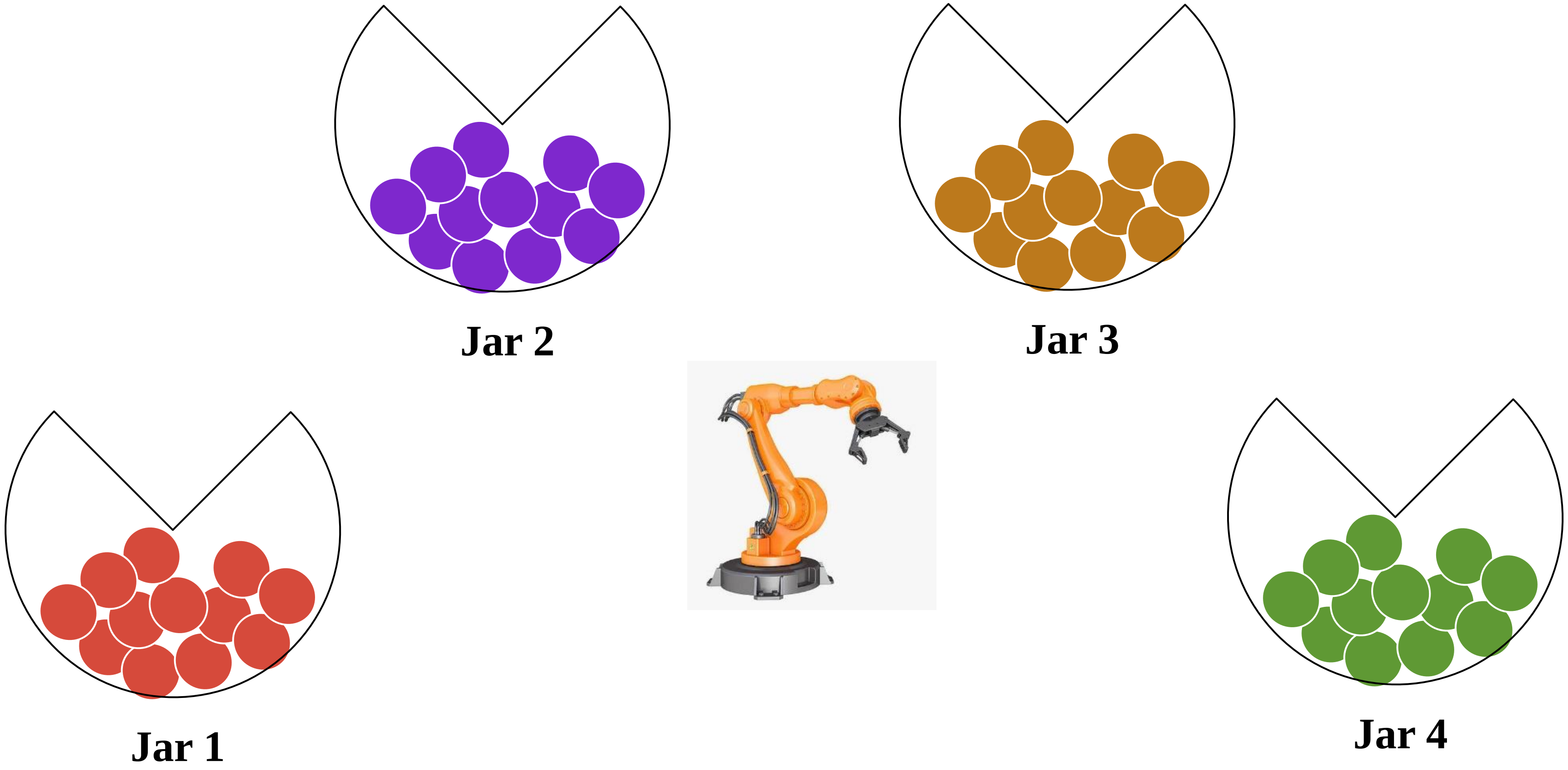
- A robot arm pick up balls from jars



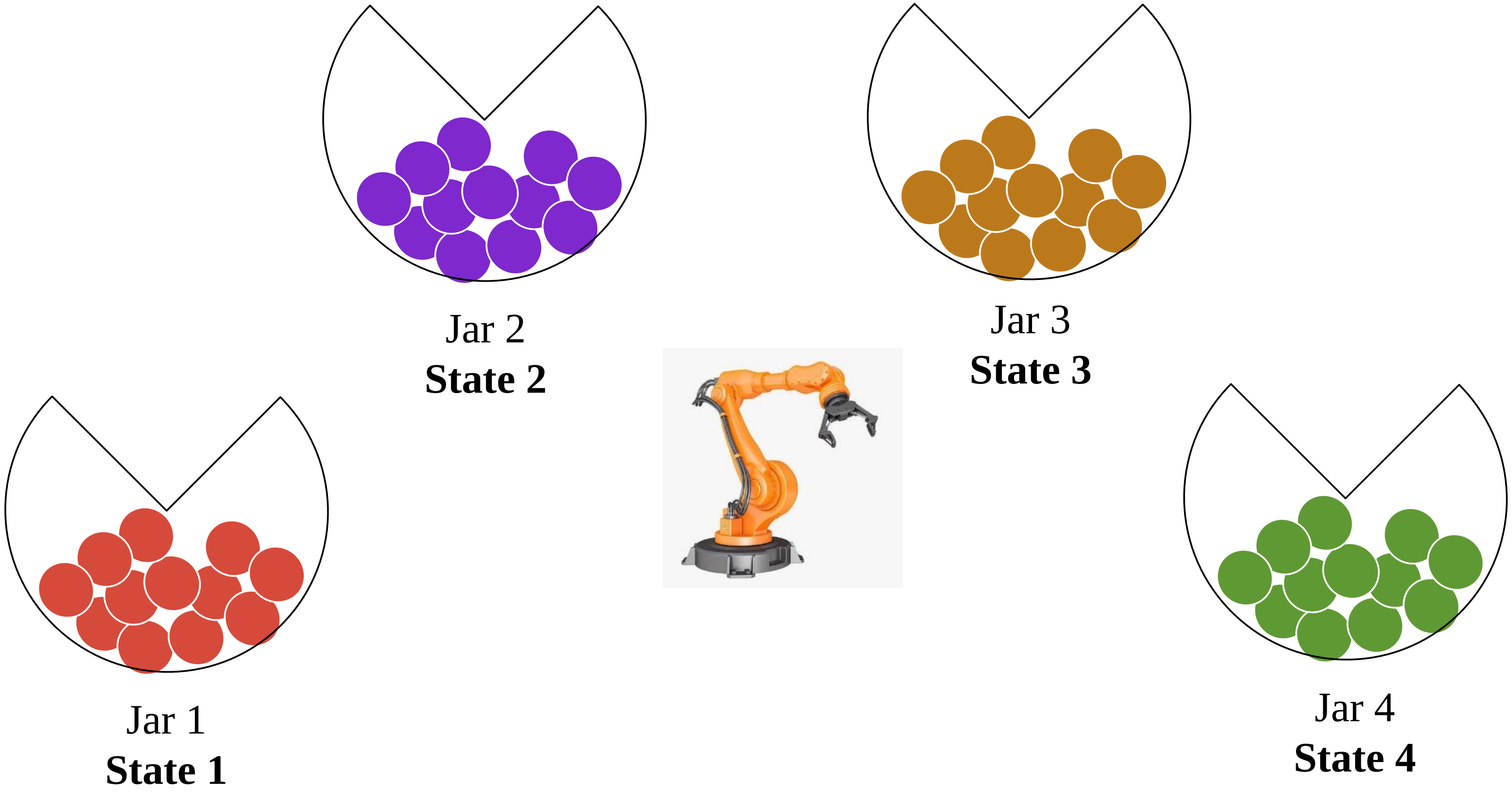
Jar 1

https://www.google.com/imgres?imgurl=https%3A%2F%2F5.imimg.com%2Fdata5%2FEE%2FEM%2FMY-4370182%2Frobot-arm-500x500.jpg&imgrefurl=https%3A%2F%2Fwww.indiamart.com%2Fproddetail%2Findustrial-robot-arm-18072085973.html&tbnid=kEmmq4Dozcd6zM&vet=12ahUKEwjauvvcy4f6AhU2gGMGHZd7A5wQMygCegUIARDvAQ..i&docid=aCDlxZsII6s0RM&w=500&h=500&q=robotic%20arm&ved=2ahUKEwjauvvcy4f6AhU2gGMGHZd7A5wQMygCegUIARDvAQ

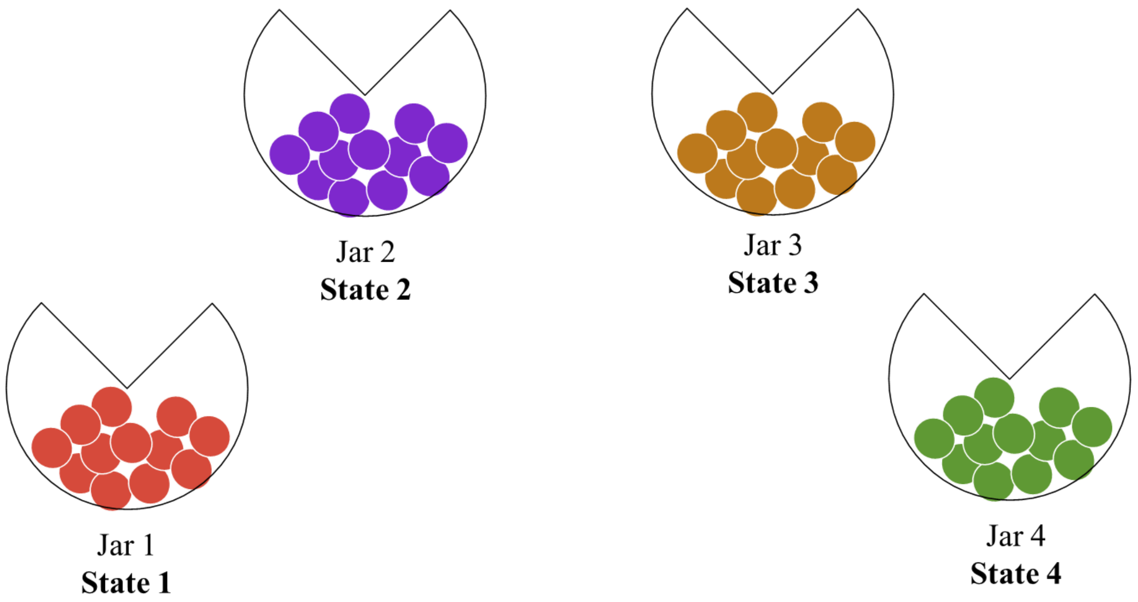
Markov Model with an Example



Markov Model with an Example

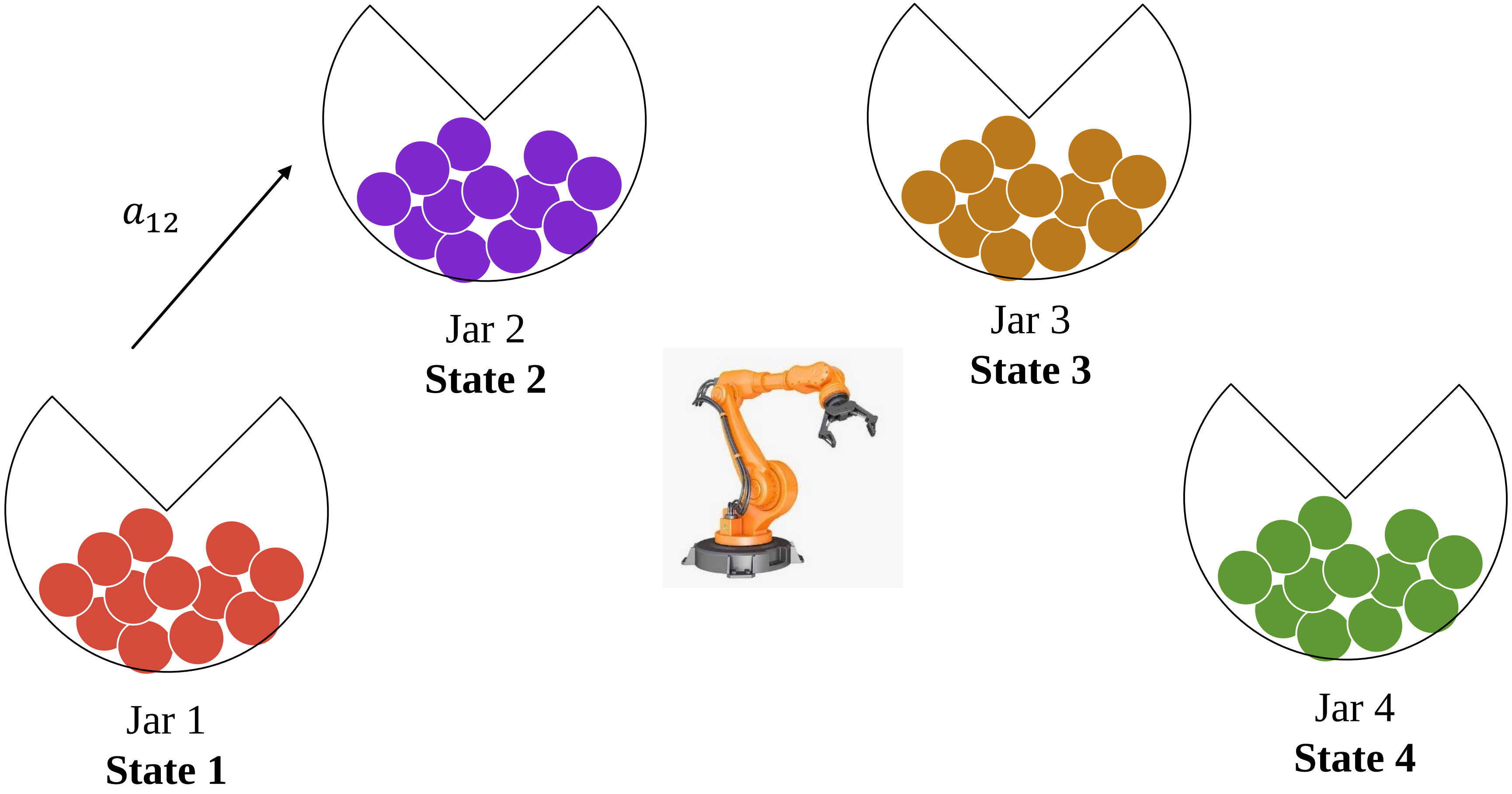


Markov Model with an Example

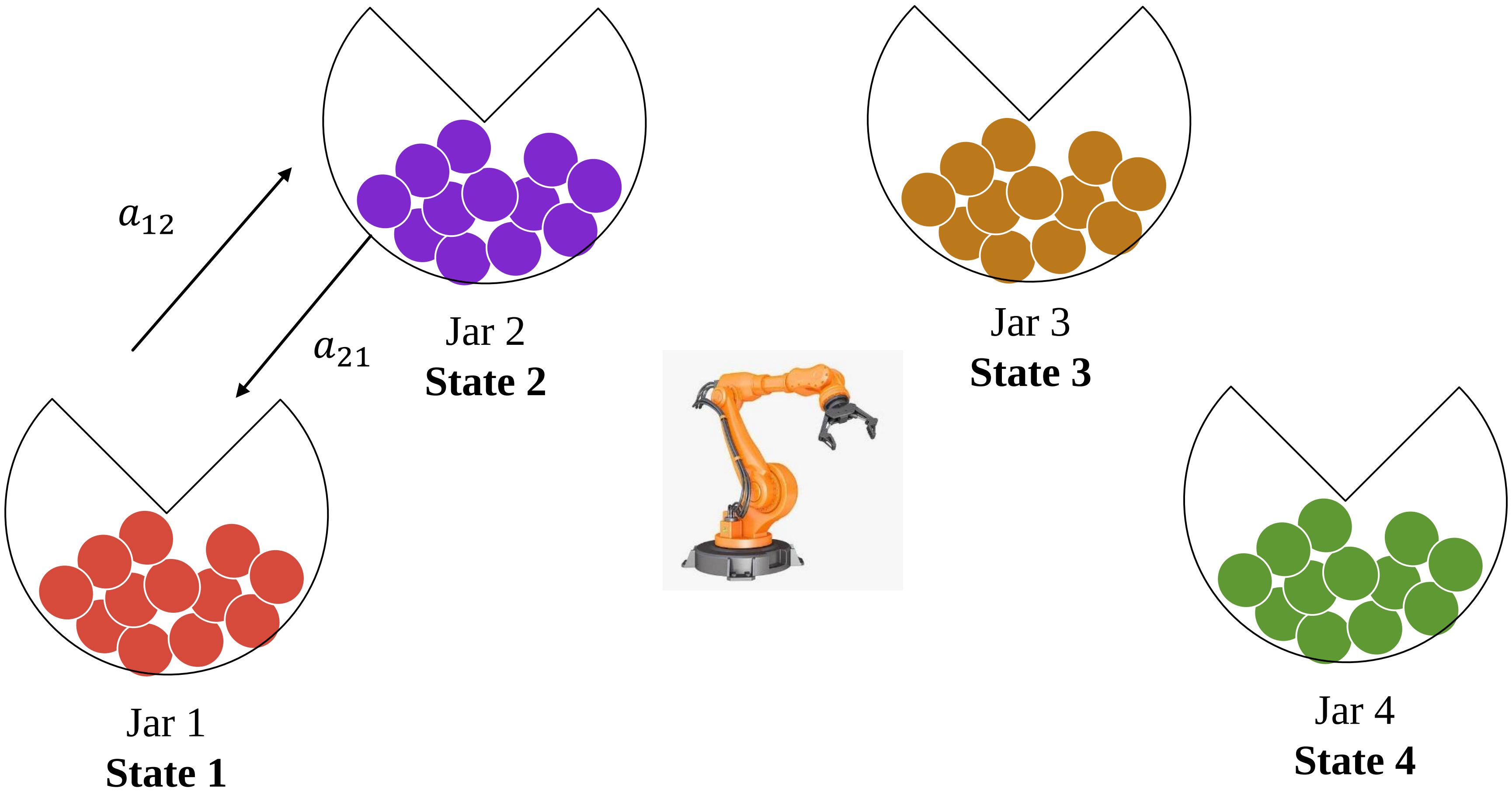


- We have a set of states (jars) and an observation (colour of the ball picked up) at each state
- The next state depends only on the present state

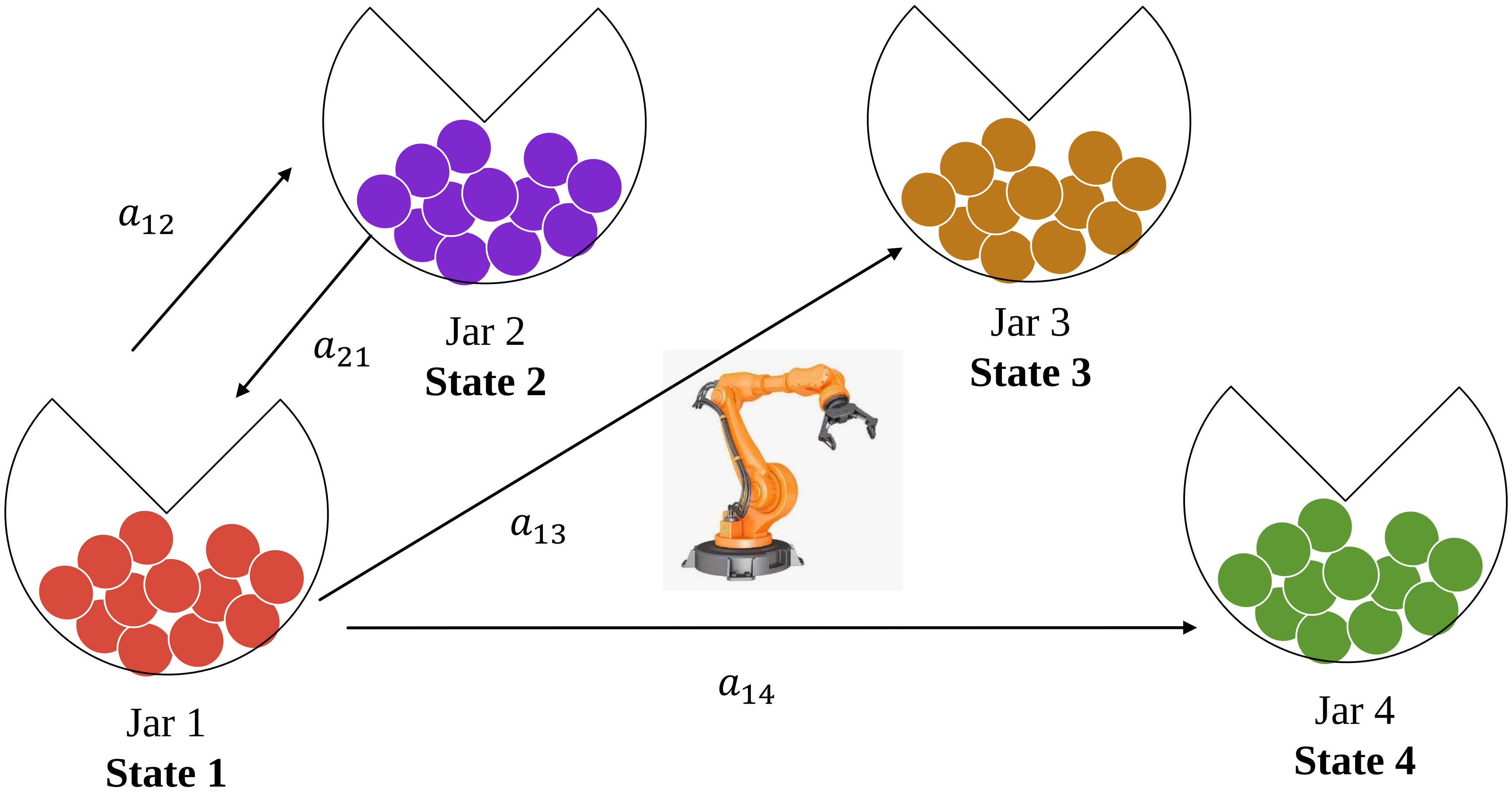
Markov Model with an Example



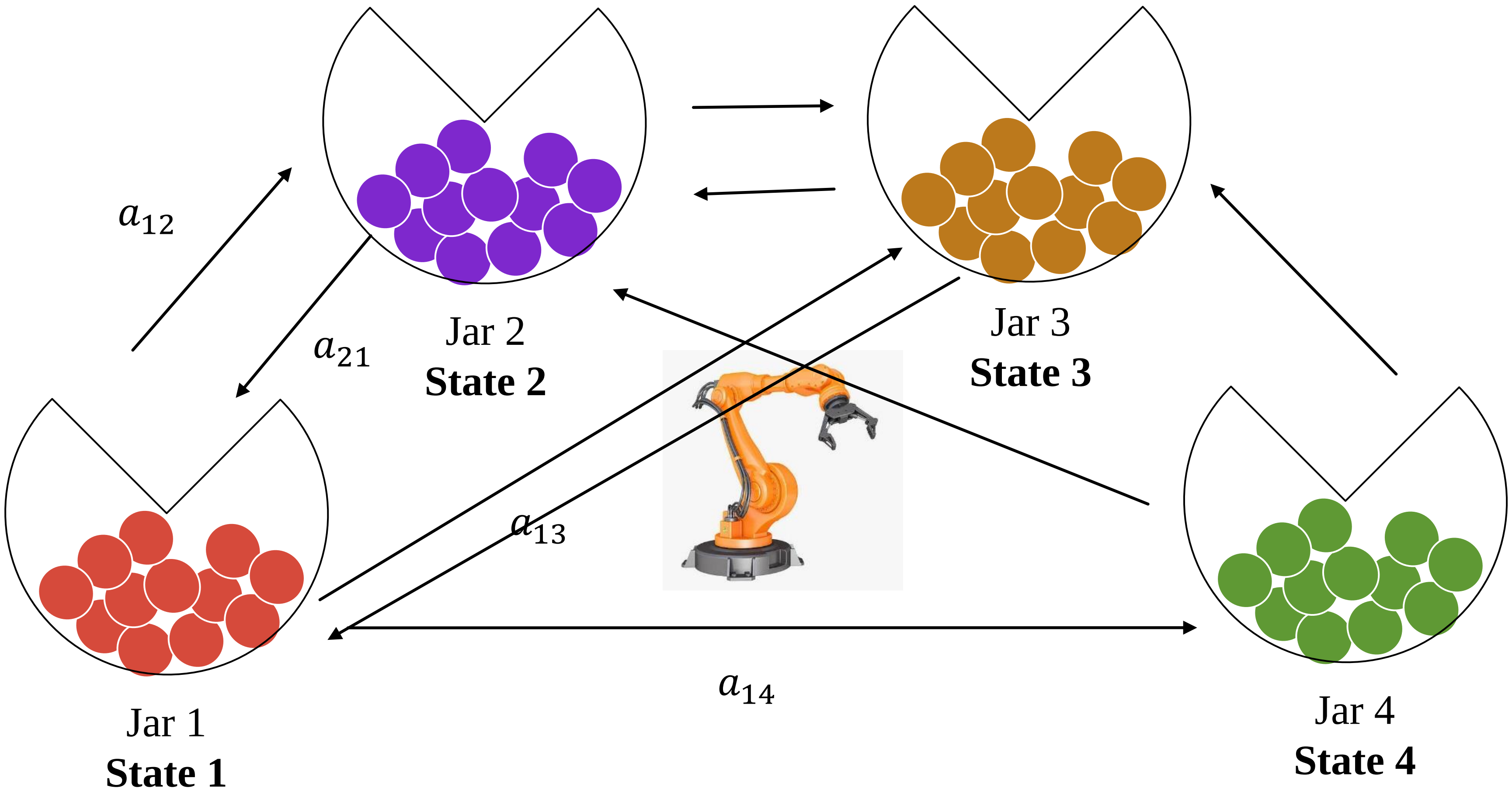
Markov Model with an Example



Markov Model with an Example

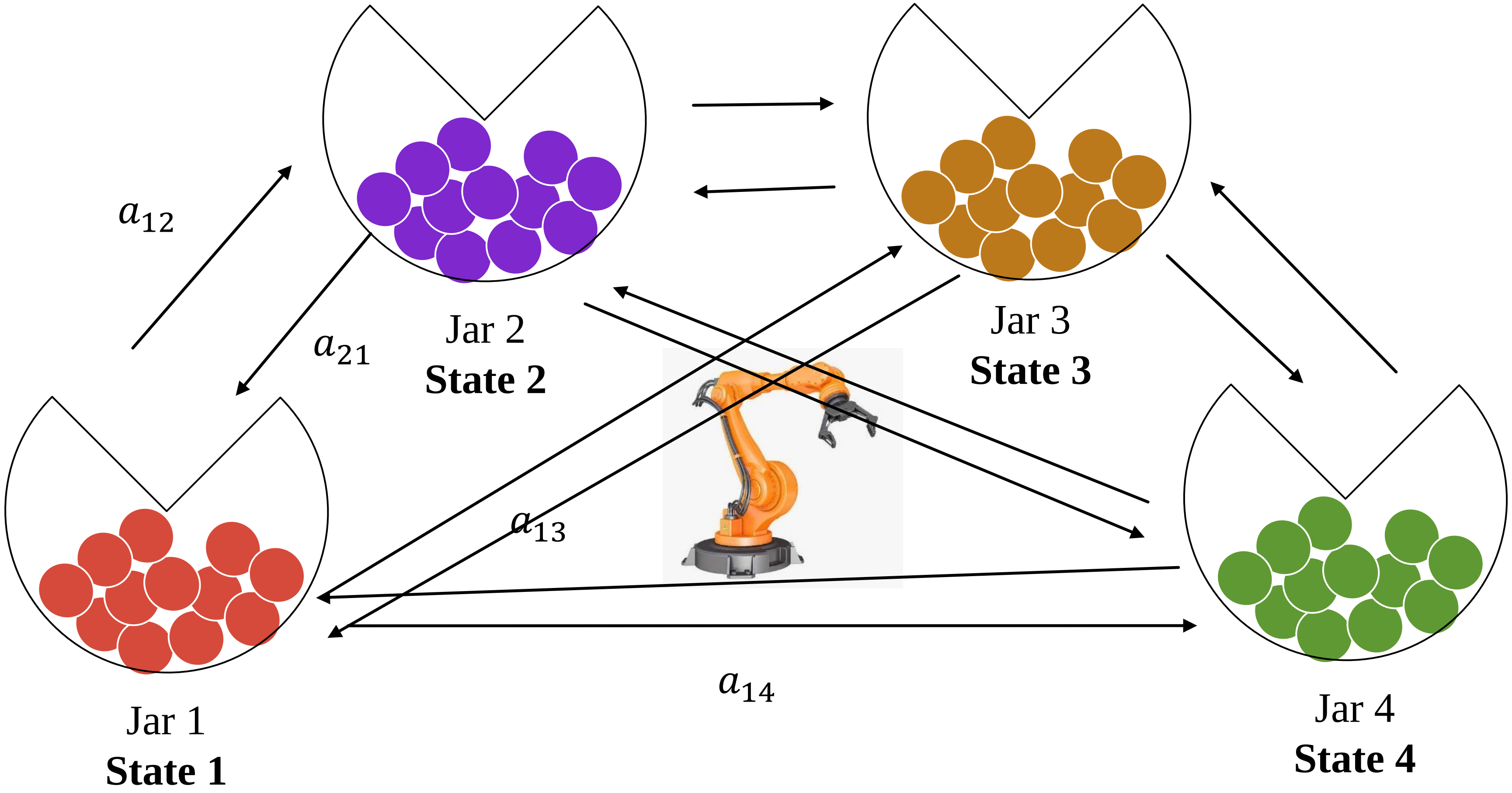


Markov Model with an Example

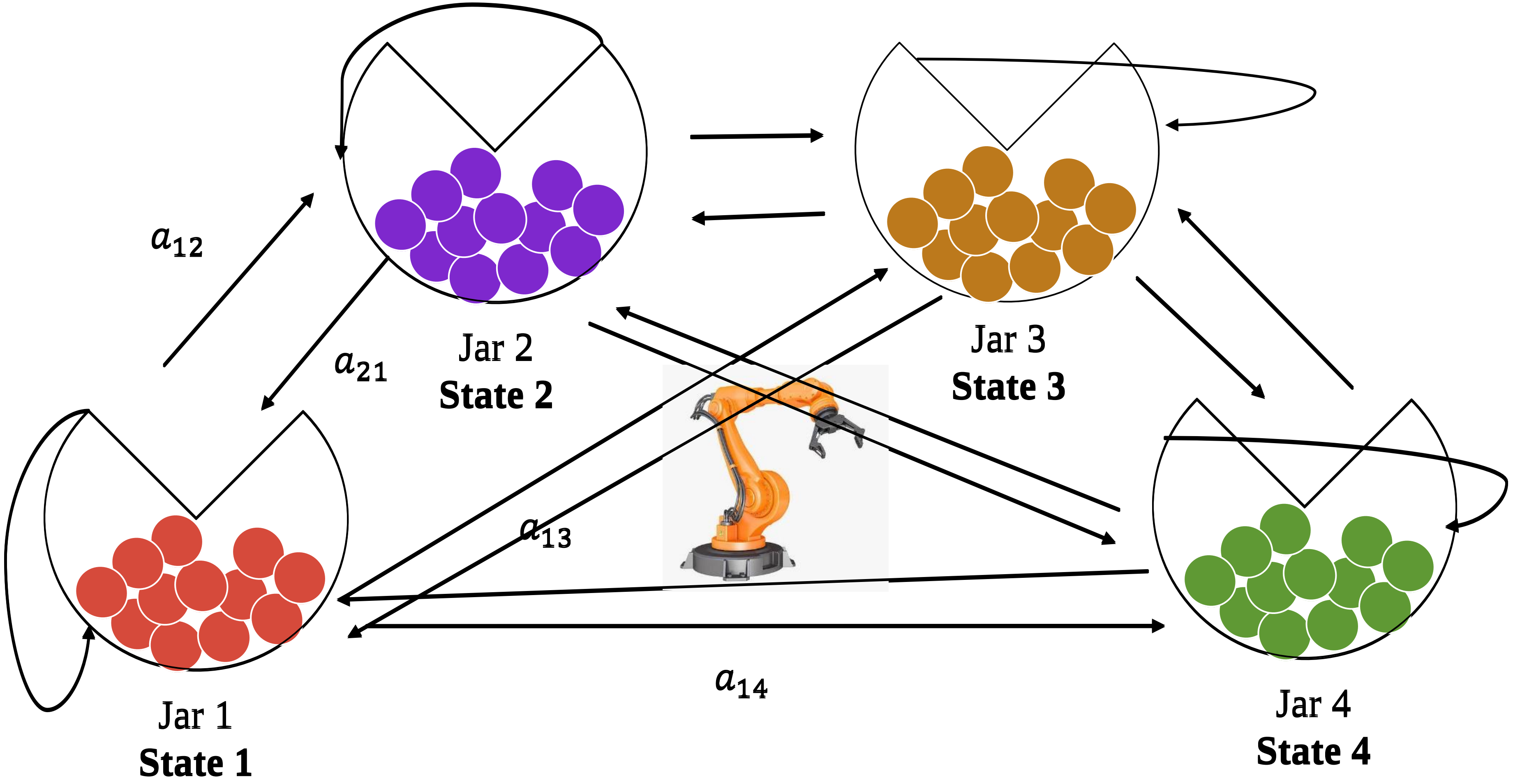


And so on ...

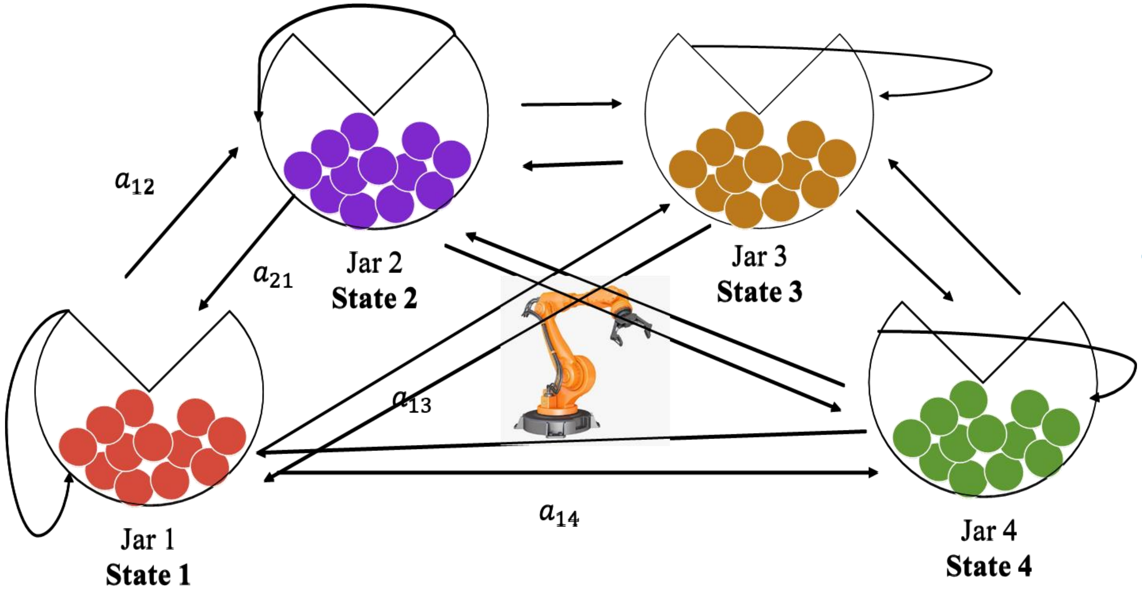
Markov Model with an Example



Markov Model with an Example

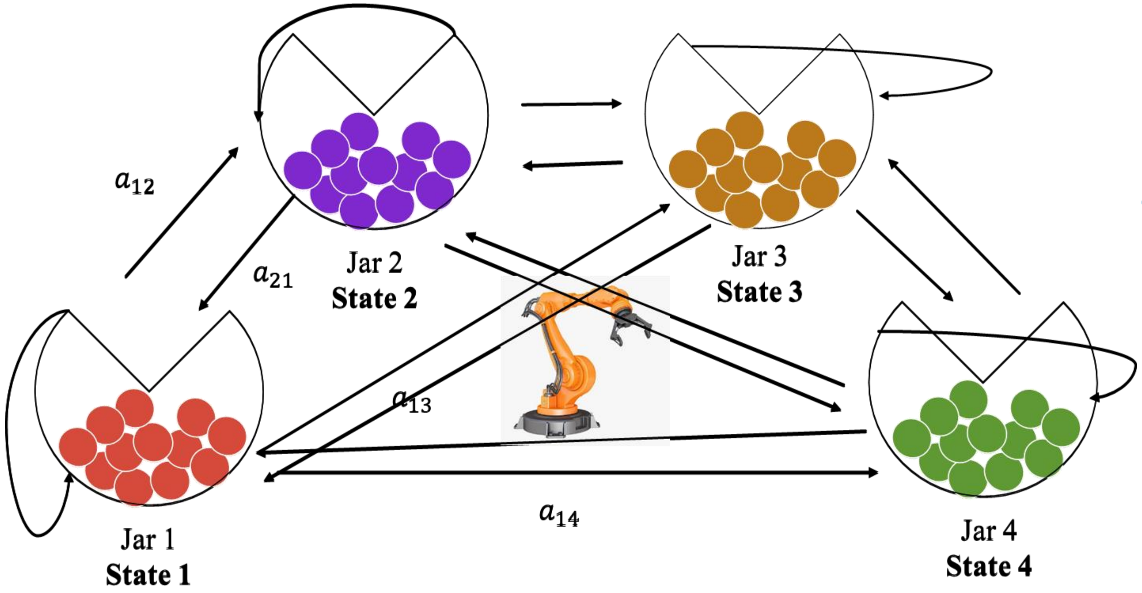


Markov Model with an Example



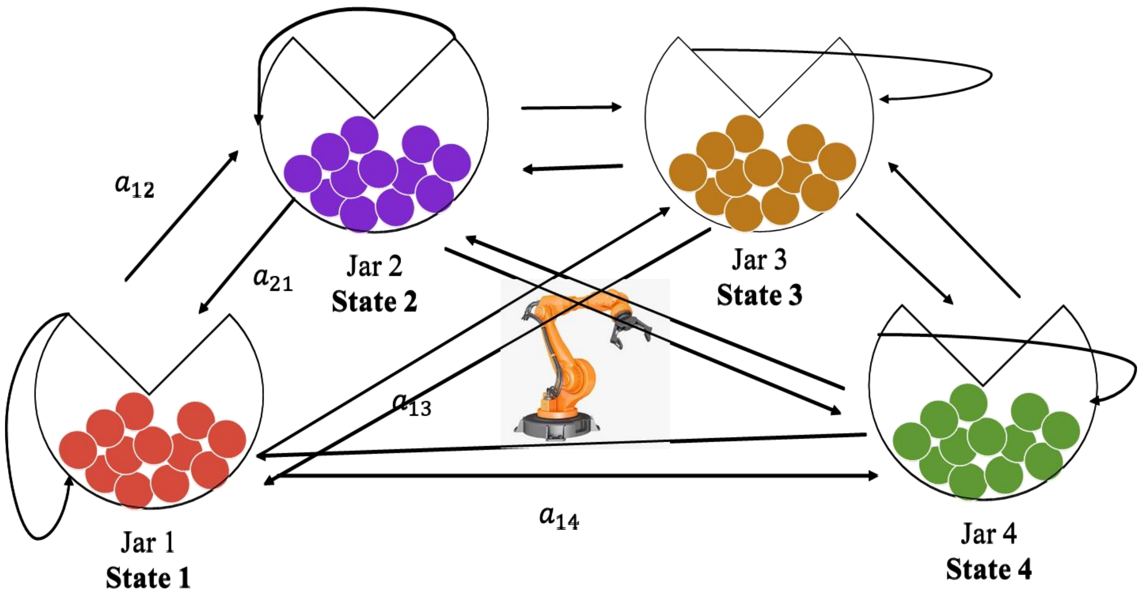
- The arm moves from one state (jar) to another state (jar) with certain probability
- However, when it reaches to a particular state our observation (the colour of the ball) is fixed
- For example, whenever the arm reach state 4, the observation is **green**

Markov Model with an Example



- Suppose, we have an observation sequence

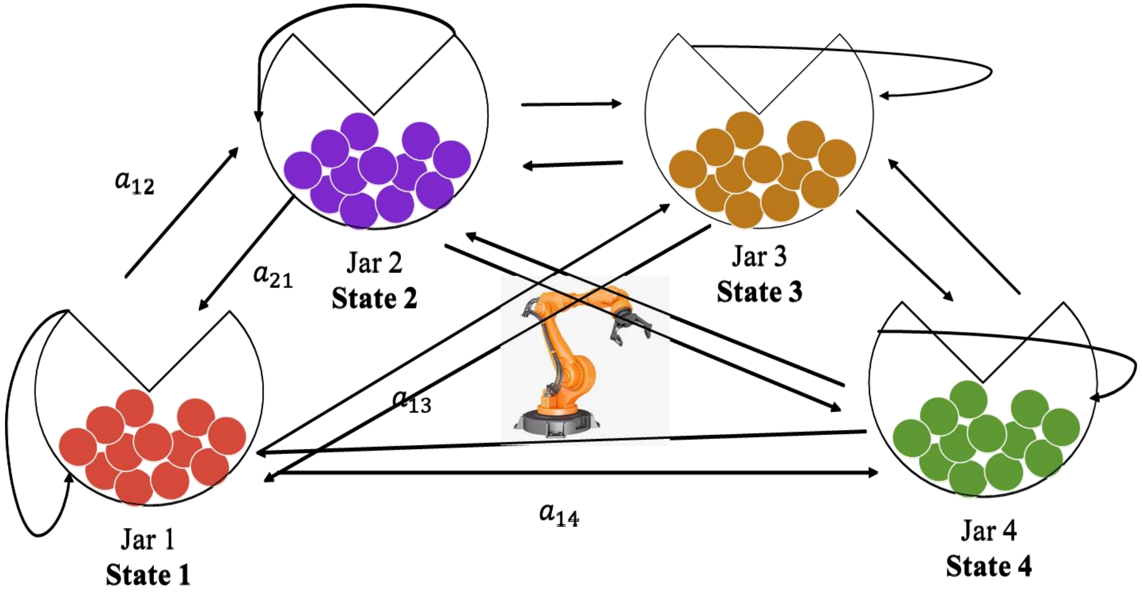
Markov Model with an Example



- Suppose, we have an observation sequence

- **What can we conclude from this?**

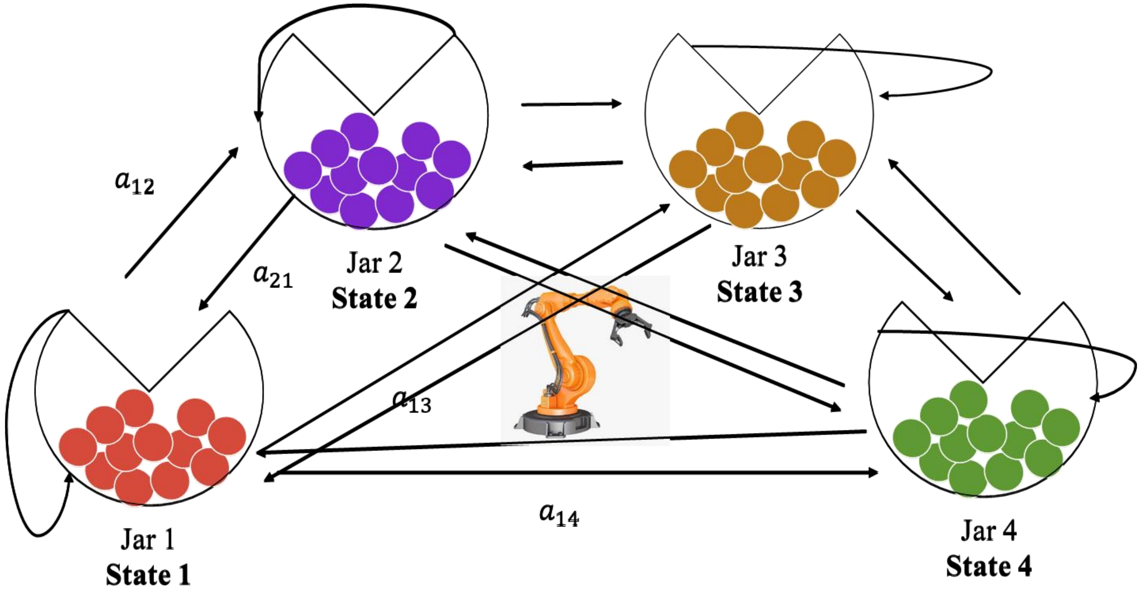
Markov Model with an Example



- Suppose, we have an observation sequence

- **What can we conclude from this?**
- **The arm started in state 2, then moved to state 3, state 3, and finally state 1**

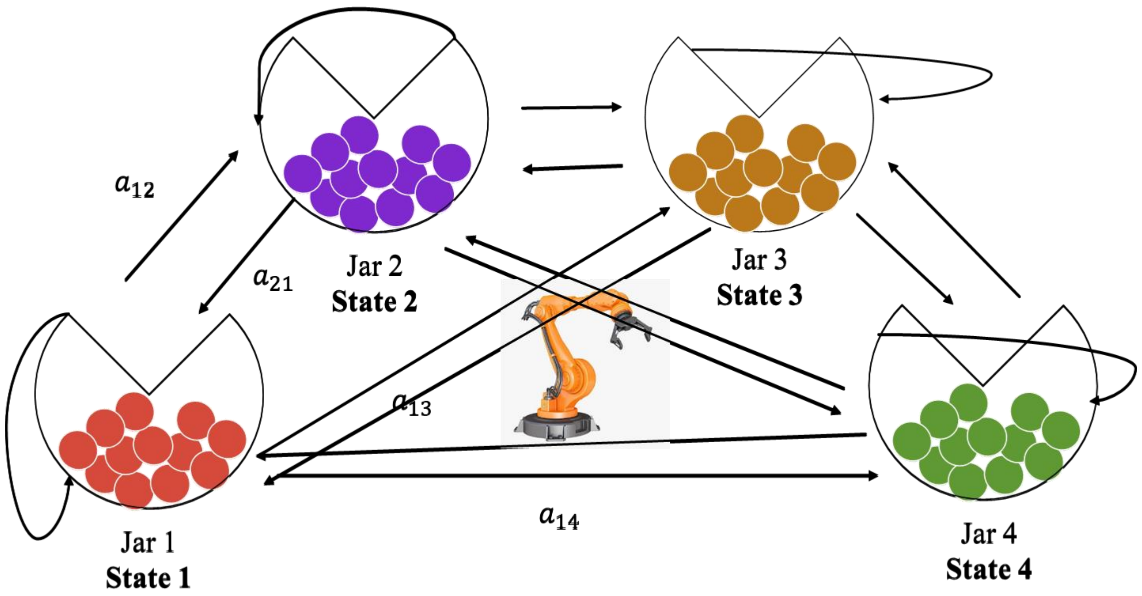
Markov Model with an Example



- Suppose, we have an observation sequence

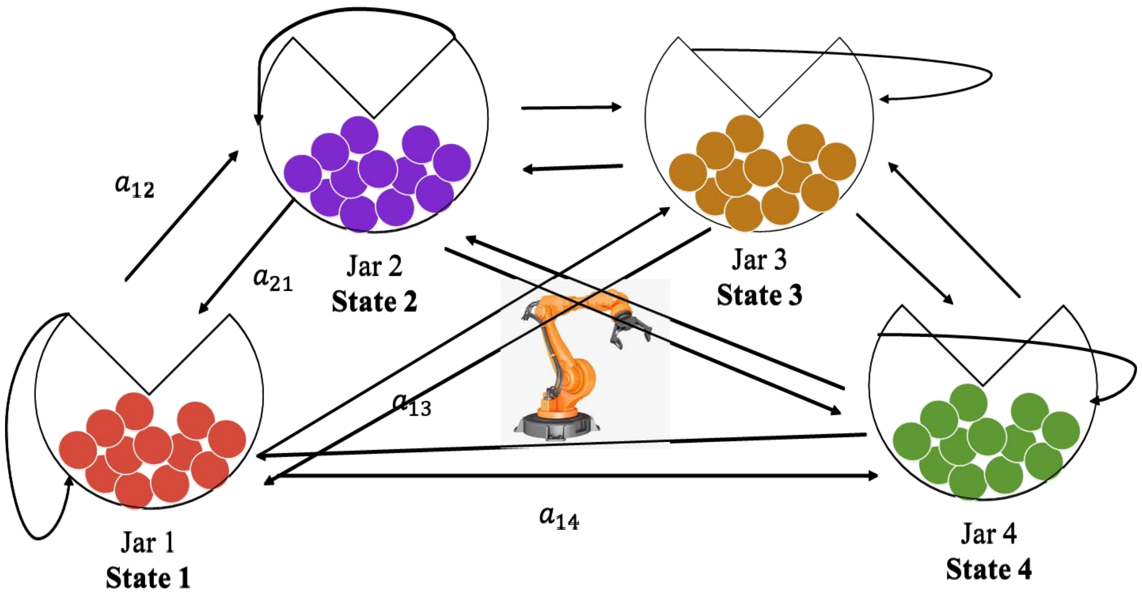
- **Observable Markov Model**

Discrete Markov Process



- State change at regularly discrete times

What Kind of Decisions Can We Make?

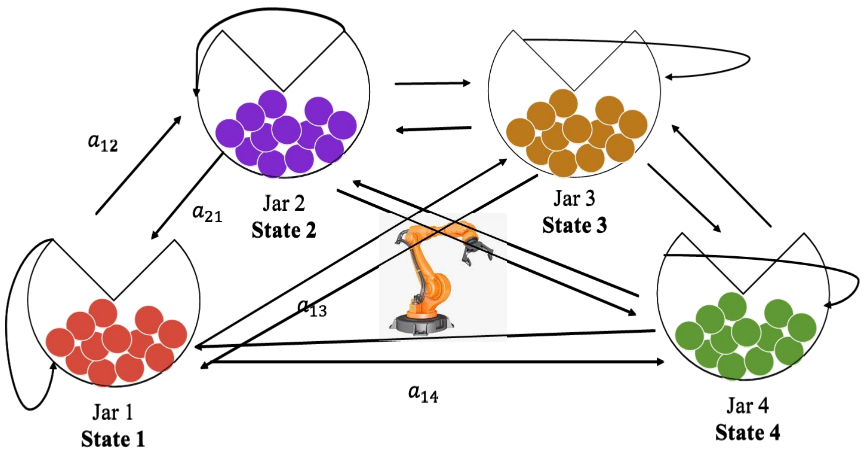


- Let's say we have an observation sequence

$$O = \{B, G, Y, Y, R, G\}$$

- What is $P(O|Model)$?

What Kind of Decisions Can We Make?



Sequence of states is
$\boldsymbol{S_2, S_4, S_3, S_3, S_1, S_4}$

$$P(O|Model) = P(B, G, Y, Y, R, G|Model)$$

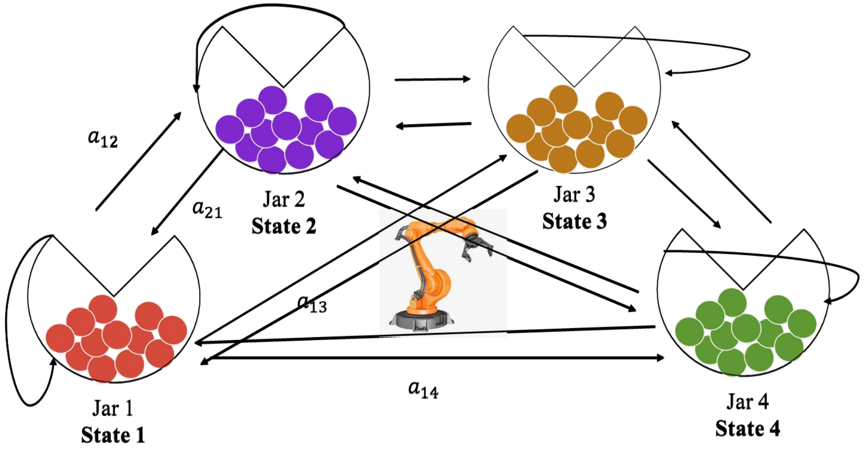
$$= P(S_2, S_4, S_3, S_3, S_1, S_4|Model)$$

$$= P(S_2|Model)P(S_4|S_2, Model)P(S_3|S_4, Model)P(S_3|S_3, Model)P(S_1|S_3, Model)P(S_4|S_1, Model)$$

$$= \pi_2 a_{24} a_{43} a_{33} a_{31} a_{14}$$

What Kind of Decisions Can We Make?

Let the model be λ



Sequence of states is
$\boldsymbol{S_2, S_4, S_3, S_3, S_1, S_4}$

$$P(O|\lambda) = P(B, G, Y, Y, R, G|\lambda)$$

$$= P(S_2, S_4, S_3, S_3, S_1, S_4|\lambda)$$

$$= P(S_2|\lambda)P(S_4|S_2, \lambda)P(S_3|S_4, \lambda)P(S_3|S_3, \lambda)P(S_1|S_3, \lambda)P(S_4|S_1, \lambda)$$

$$= \pi_2 a_{24} a_{43} a_{33} a_{31} a_{14}$$

Can We Prove?

Sequence of states is
$\boldsymbol{S_2, S_4, S_3, S_3, S_1, S_4}$

$$P(S_2, S_4, S_3, S_3, S_1, S_4|\lambda) = P(S_2|\lambda)P(S_4|S_2, \lambda)P(S_3|S_4, \lambda)P(S_3|S_3, \lambda)P(S_1|S_3, \lambda)P(S_4|S_1, \lambda)$$

Can We Prove?

Sequence of states is
$\boldsymbol{S_2, S_4, S_3, S_3, S_1, S_4}$

$$\boldsymbol{P(S_2, S_4, S_3, S_3, S_1, S_4|\lambda) = P(S_2|\lambda)P(S_4|S_2,\lambda)P(S_3|S_4,\lambda)P(S_3|S_3,\lambda)P(S_1|S_3,\lambda)P(S_4|S_1,\lambda)}$$

$$P(S_2, S_4, S_3, S_3, S_1, S_4, \lambda) = P(S_2, S_4, S_3, S_3, S_1, S_4|\lambda)P(\lambda)$$

$$P(S_2, S_4, S_3, S_3, S_1, S_4|\lambda) = \frac{P(S_2, S_4, S_3, S_3, S_1, S_4, \lambda)}{P(\lambda)}$$

Sequence of states is
$\boldsymbol{S_2, S_4, S_3, S_3, S_1, S_4}$

Can We Prove?

$$\boldsymbol{P(S_2, S_4, S_3, S_3, S_1, S_4|\lambda) = P(S_2|\lambda)P(S_4|S_2, \lambda)P(S_3|S_4, \lambda)P(S_3|S_3, \lambda)P(S_1|S_3, \lambda)P(S_4|S_1, \lambda)}$$

$$P(S_2, S_4, S_3, S_3, S_1, S_4, \lambda) = P(S_4|S_2, S_4, S_3, S_3, S_1, \lambda)P(S_2, S_4, S_3, S_3, S_1, \lambda)$$

$$= P(S_4|S_2, S_4, S_3, S_3, S_1, \lambda)P(S_1|S_2, S_4, S_3, S_3, \lambda)P(S_2, S_4, S_3, S_3, \lambda)$$

$$= P(S_4|S_2, S_4, S_3, S_3, S_1, \lambda)P(S_1|S_2, S_4, S_3, S_3, \lambda)P(S_3|S_2, S_4, S_3, \lambda)P(S_2, S_4, S_3, \lambda)$$

$$= P(S_4|S_2, S_4, S_3, S_3, S_1, \lambda)P(S_1|S_2, S_4, S_3, S_3, \lambda)P(S_3|S_2, S_4, S_3, \lambda)P(S_3|S_2, S_4, \lambda)P(S_2, S_4, \lambda)$$

$$= P(S_4|S_2, S_4, S_3, S_3, S_1, \lambda)P(S_1|S_2, S_4, S_3, S_3, \lambda)P(S_3|S_2, S_4, S_3, \lambda)P(S_3|S_2, S_4, \lambda)P(S_4|S_2, \lambda)P(S_2, \lambda)$$

$$= P(S_4|S_2, S_4, S_3, S_3, S_1, \lambda)P(S_1|S_2, S_4, S_3, S_3, \lambda)P(S_3|S_2, S_4, S_3, \lambda)P(S_3|S_2, S_4, \lambda)P(S_4|S_2, \lambda)P(S_2|\lambda)P(\lambda)$$

Sequence of states is
$\boldsymbol{S_2, S_4, S_3, S_3, S_1, S_4}$

Can We Prove?

$$\boldsymbol{P(S_2, S_4, S_3, S_3, S_1, S_4|\lambda) = P(S_2|\lambda)P(S_4|S_2, \lambda)P(S_3|S_4, \lambda)P(S_3|S_3, \lambda)P(S_1|S_3, \lambda)P(S_4|S_1, \lambda)}$$

$$P(S_2, S_4, S_3, S_3, S_1, S_4, \lambda)$$
$$= P(S_4|S_2, S_4, S_3, S_3, S_1, \lambda)P(S_1|S_2, S_4, S_3, S_3, \lambda)P(S_3|S_2, S_4, S_3, \lambda)P(S_3|S_2, S_4, \lambda)P(S_4|S_2, \lambda)P(S_2|\lambda)P(\lambda)$$

Sequence of states is
$\boldsymbol{S_2, S_4, S_3, S_3, S_1, S_4}$

Can We Prove?

$$\boldsymbol{P(S_2, S_4, S_3, S_3, S_1, S_4|\lambda) = P(S_2|\lambda)P(S_4|S_2, \lambda)P(S_3|S_4, \lambda)P(S_3|S_3, \lambda)P(S_1|S_3, \lambda)P(S_4|S_1, \lambda}$$

$$P(S_2, S_4, S_3, S_3, S_1, S_4, \lambda)$$
$$= P(S_4|S_2, S_4, S_3, S_3, S_1, \lambda)P(S_1|S_2, S_4, S_3, S_3, \lambda)P(S_3|S_2, S_4, S_3, \lambda)P(S_3|S_2, S_4, \lambda)P(S_4|S_2, \lambda)P(S_2|\lambda)P(\lambda)$$

$$= P(S_4|S_1, \lambda)P(S_1|S_3, \lambda)P(S_3|S_3, \lambda)P(S_3|S_4, \lambda)P(S_4|S_2, \lambda)P(S_2|\lambda)P(\lambda)$$

Sequence of states is $S_2, S_4, S_3, S_3, S_1, S_4$

Can We Prove?

$$\boldsymbol{P(S_2, S_4, S_3, S_3, S_1, S_4|\lambda) = P(S_2|\lambda)P(S_4|S_2,\lambda)P(S_3|S_4,\lambda)P(S_3|S_3,\lambda)P(S_1|S_3,\lambda)P(S_4|S_1,\lambda)}$$

$$P(S_2, S_4, S_3, S_3, S_1, S_4, \lambda)$$
$$= P(S_4|S_1,\lambda)P(S_1|S_3,\lambda)P(S_3|S_3,\lambda)P(S_3|S_4,\lambda)P(S_4|S_2,\lambda)P(S_2|\lambda)P(\lambda)$$

$$\boldsymbol{P(S_2, S_4, S_3, S_3, S_1, S_4|\lambda) = \frac{P(S_2, S_4, S_3, S_3, S_1, S_4, \lambda)}{P(\lambda)}}$$

Can We Prove?

Sequence of states is $\boldsymbol{S_2, S_4, S_3, S_3, S_1, S_4}$

$$\boldsymbol{P(S_2, S_4, S_3, S_3, S_1, S_4|\lambda) = P(S_2|\lambda)P(S_4|S_2,\lambda)P(S_3|S_4,\lambda)P(S_3|S_3,\lambda)P(S_1|S_3,\lambda)P(S_4|S_1,\lambda}$$

$$P(S_2, S_4, S_3, S_3, S_1, S_4, \lambda)$$
$$= P(S_4|S_1,\lambda)P(S_1|S_3,\lambda)P(S_3|S_3,\lambda)P(S_3|S_4,\lambda)P(S_4|S_2,\lambda)P(S_2|\lambda)P(\lambda)$$

$$\boldsymbol{P(S_2, S_4, S_3, S_3, S_1, S_4|\lambda) = \frac{P(S_4|S_1,\lambda)P(S_1|S_3,\lambda)P(S_3|S_3,\lambda)P(S_3|S_4,\lambda)P(S_4|S_2,\lambda)P(S_2|\lambda)P(\lambda)}{P(\lambda)}}$$

Can We Prove?

Sequence of states is
$\boldsymbol{S_2, S_4, S_3, S_3, S_1, S_4}$

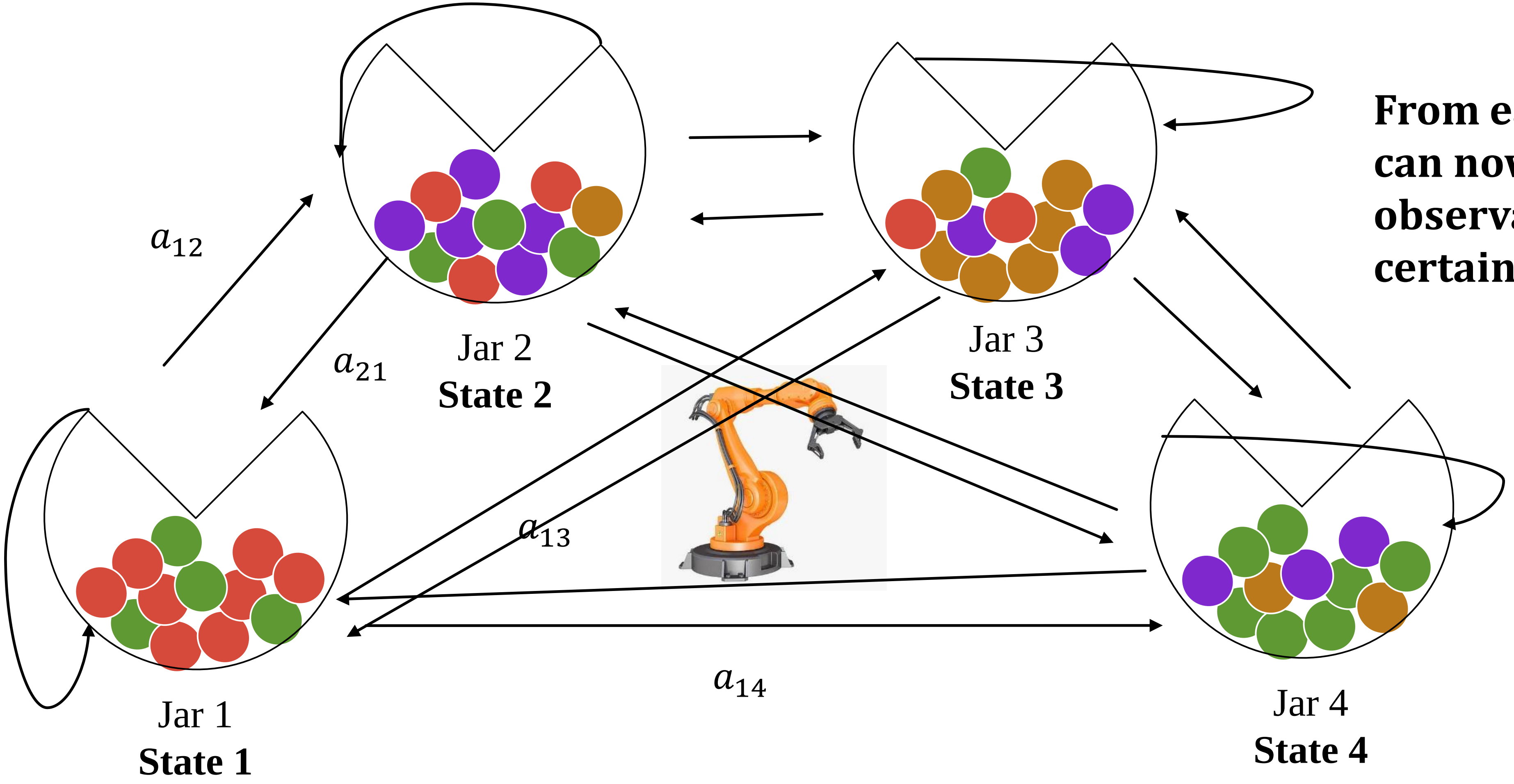
$$\boldsymbol{P(S_2, S_4, S_3, S_3, S_1, S_4|\lambda) = P(S_2|\lambda)P(S_4|S_2, \lambda)P(S_3|S_4, \lambda)P(S_3|S_3, \lambda)P(S_1|S_3, \lambda)P(S_4|S_1, \lambda}$$

$$P(S_2, S_4, S_3, S_3, S_1, S_4, \lambda)$$
$$= P(S_4|S_1, \lambda)P(S_1|S_3, \lambda)P(S_3|S_3, \lambda)P(S_3|S_4, \lambda)P(S_4|S_2, \lambda)P(S_2|\lambda)P(\lambda)$$

$$\boldsymbol{P(S_2, S_4, S_3, S_3, S_1, S_4|\lambda) = P(S_4|S_1, \lambda)P(S_1|S_3, \lambda)P(S_3|S_3, \lambda)P(S_3|S_4, \lambda)P(S_4|S_2, \lambda)P(S_2|\lambda)}$$

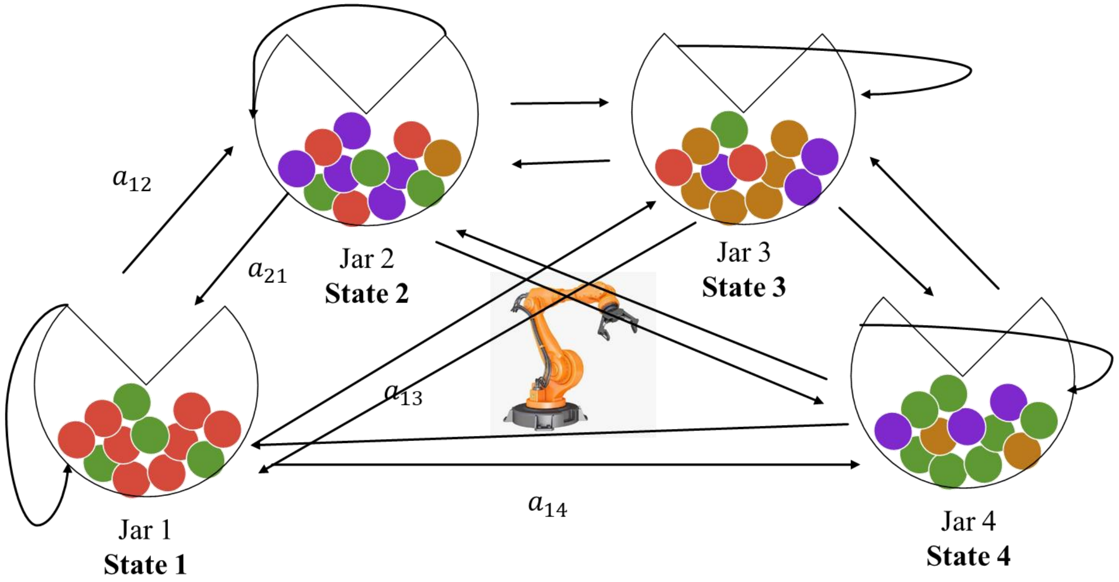
$$\boldsymbol{= P(S_2|\lambda)P(S_4|S_2, \lambda)P(S_3|S_4, \lambda)P(S_3|S_3, \lambda)P(S_1|S_3, \lambda)P(S_4|S_1, \lambda)}$$

Now, Consider This Situation



From each state, we can now have multiple observations with certain probabilities

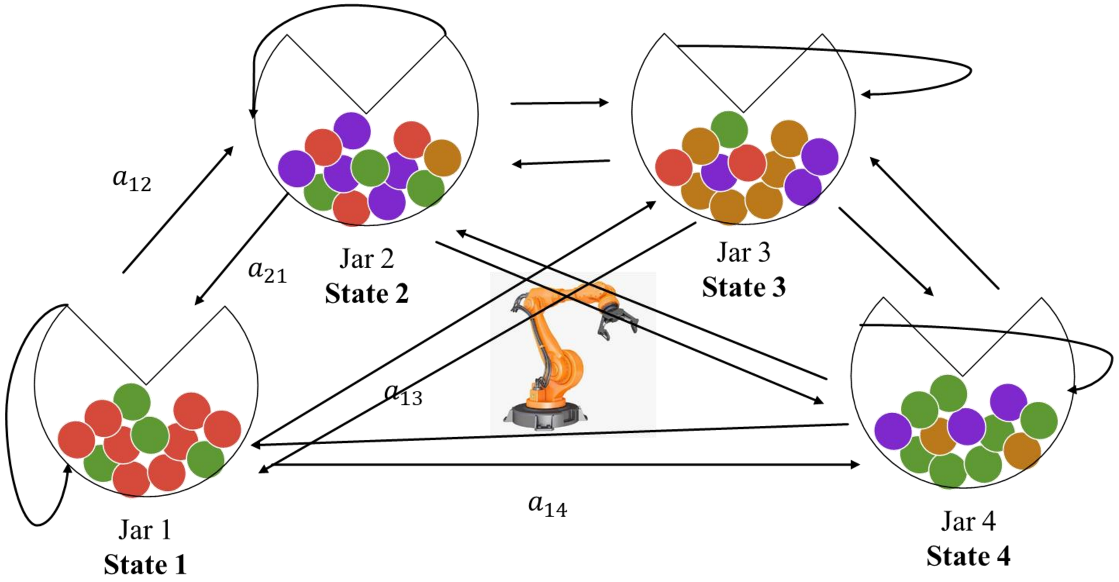
The New Situation



- Suppose, we have an observation sequence

- **What can we conclude from this?**

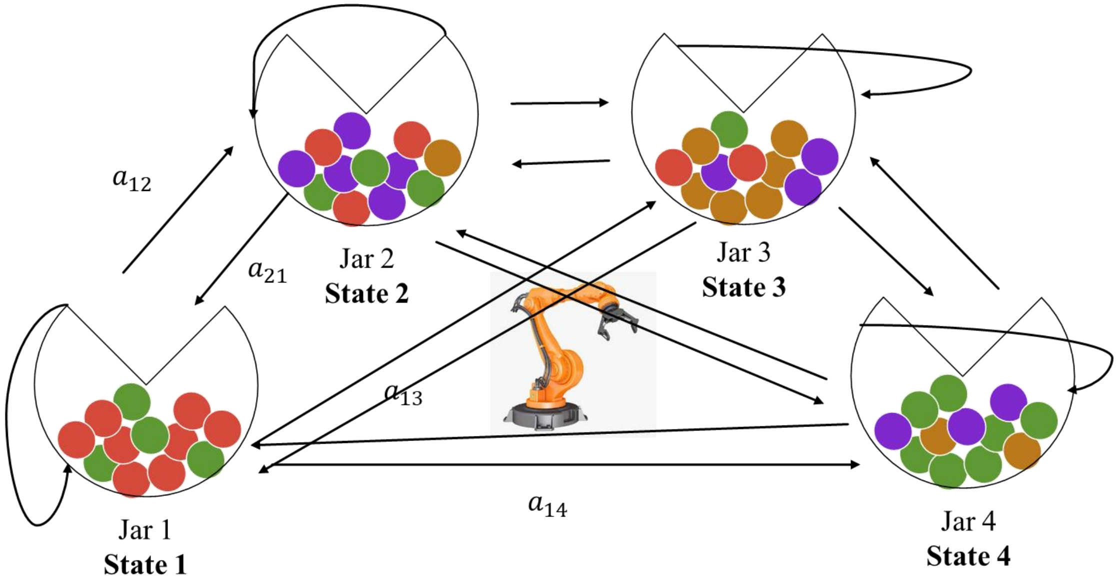
The New Situation



- Suppose, we have an observation sequence

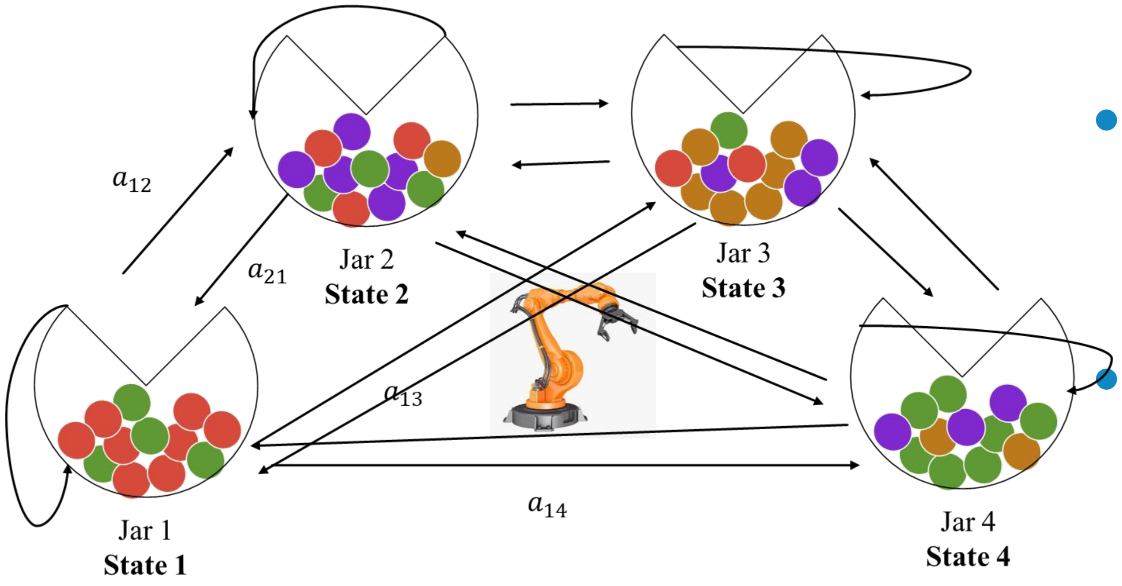
- **What can we conclude from this?**
- **We can't conclude about the movement of the arm**
- **we can only make probabilistic decisions about the movement through the states (jars)**

The Hidden Markov Model



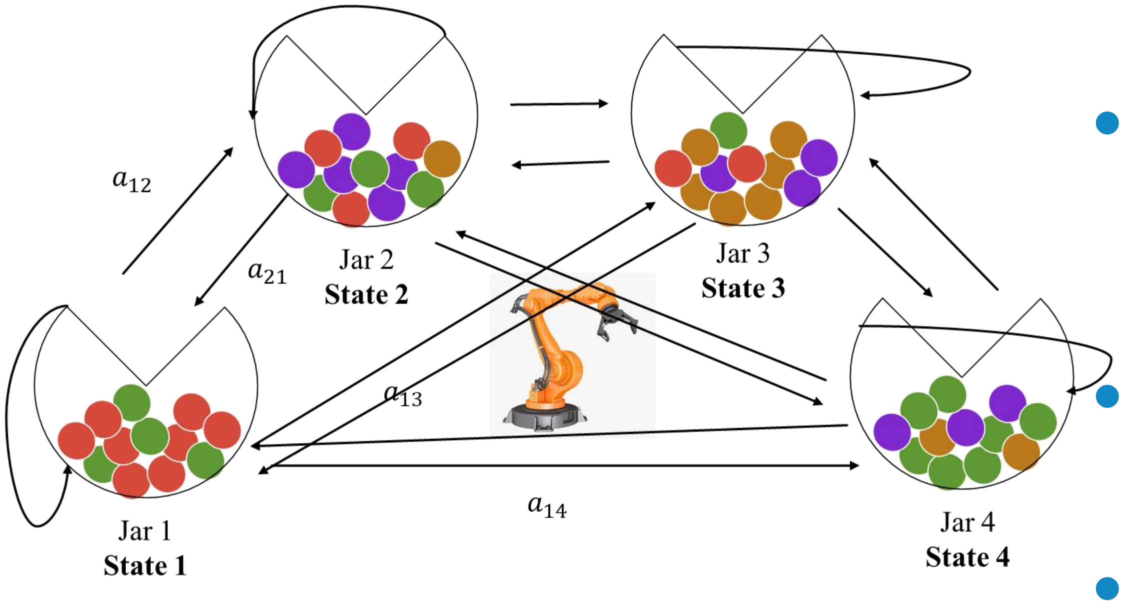
- The states (jars) traversed during the movement are hidden (not known to us)
- We have a set of observations

The Hidden Markov Model



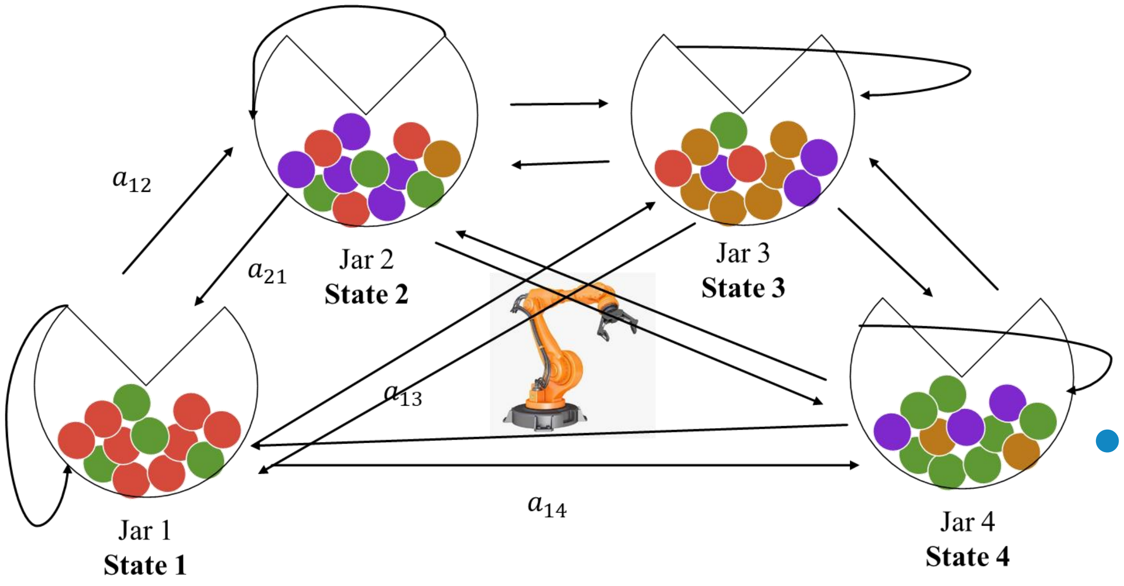
- A set of states S
  - N states
- A set of observations O
  - At each state, we may have M different observations
  - $v_1, v_2, \ldots, v_M$
- State transition matrix A
  - State transition probabilities a_{ij}
  - $a_{ij} = P(q_t = S_j | q_{t-1} = S_i) \quad 1 \leq i, j \leq N$
  - $\sum_{j=1}^{N} a_{ij} = 1$

The Hidden Markov Model



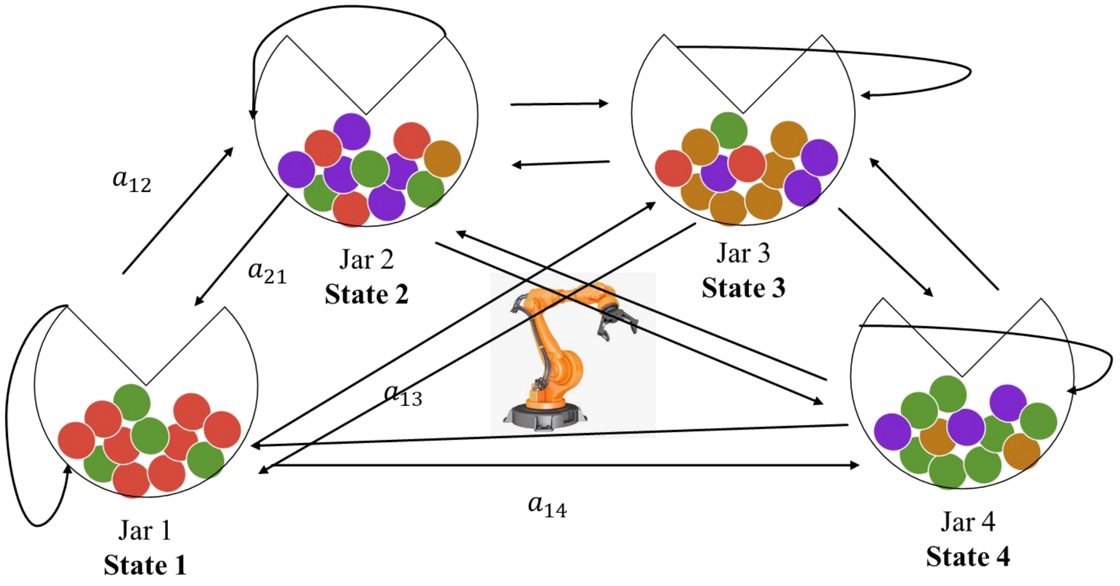
- A set of states S
  - N states
- A set of observations O
  - At each state, we may have M different observations
  - $v_1, v_2, \dots, v_M$
- State transition matrix A
  - State transition probabilities a_{ij}
- States are ergodic
  - From one state, we can move to any other state

The Hidden Markov Model



- Observation probabilities
  - $B = \{b_j(k)\}$
  - At each state, we may have M different observations/ symbols
  - $v_1, v_2, \dots, v_M$
- $b_j(k) = P(v_k \; at \; t | q_t = S_j)$
  - $1 \leq j \leq N$
  - $1 \leq k \leq M$

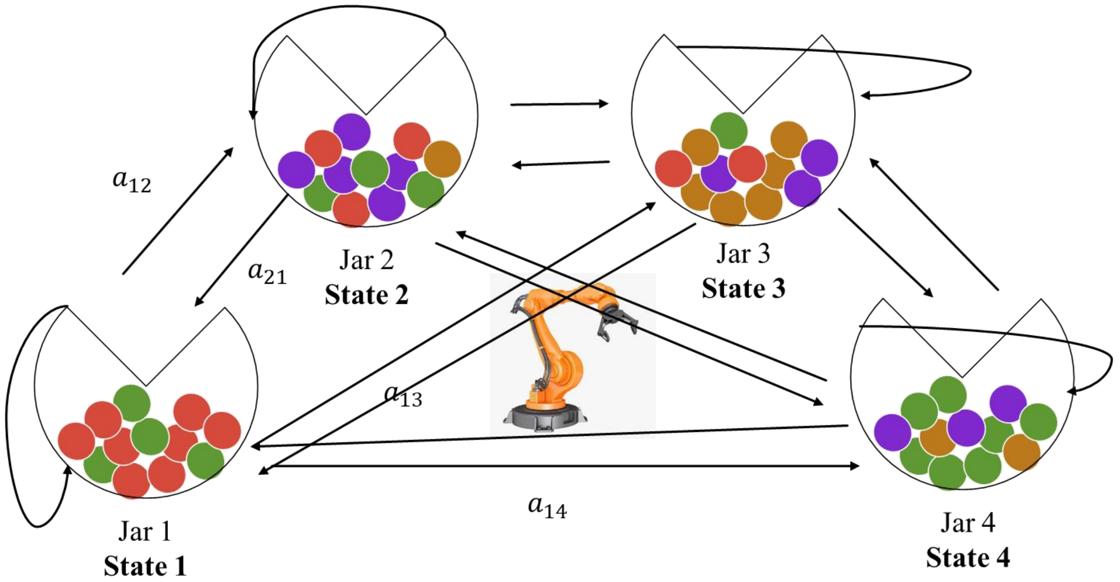
The Hidden Markov Model



- Initial state probabilities $\pi = \{\pi_i\}$

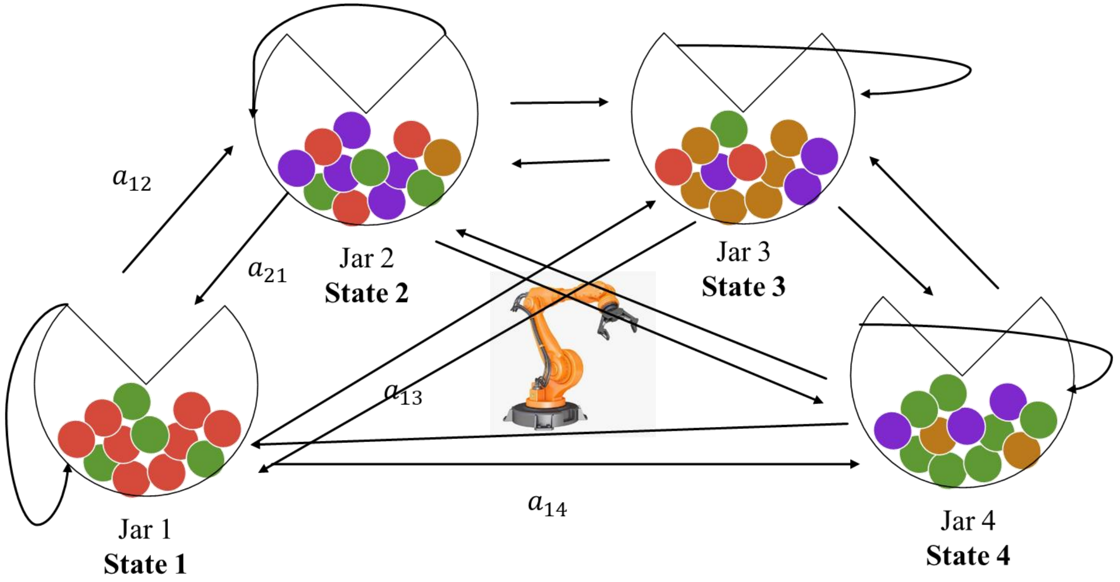
- $\pi_i = P(q_1 = S_i) \qquad 1 \le i \le N$

The Hidden Markov Model



- $\text{HMM} = \{N, M, A, B, \pi\}$
- $O = \{O_1, O_2, \ldots, O_T\}$

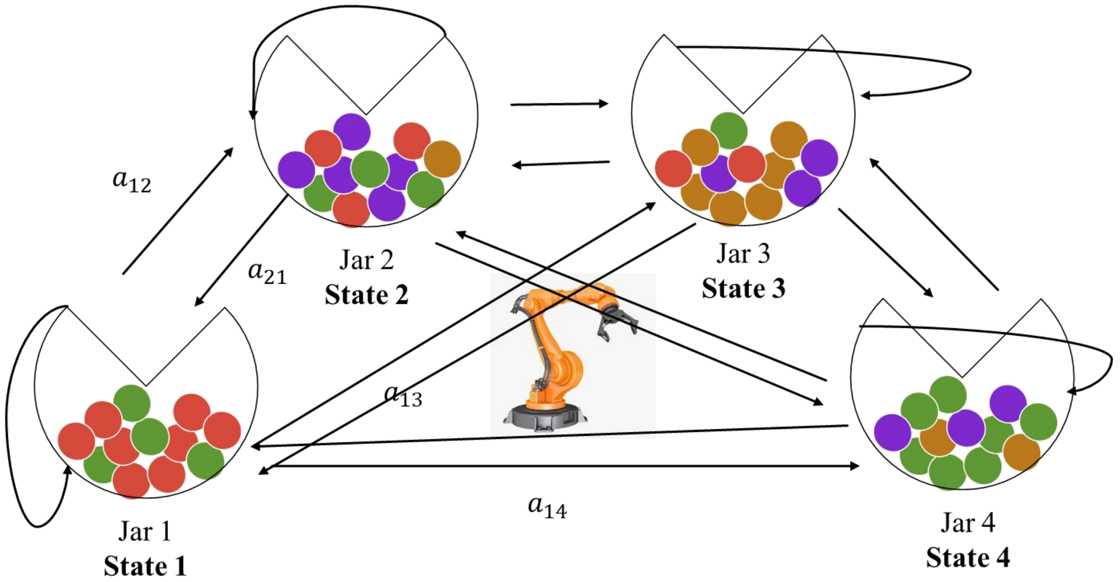
The Hidden Markov Model



- HMM $= \{N, M, A, B, \pi\}$
- $O = \{O_1, O_2, \ldots, O_T\}$

- To generate observations
  - Choose $q_1 = S_i$ according to π_i
  - Set $t = 1$
  - Choose $O_t = v_k$ according to B i.e. $b_i(k)$
  - Transit to a new state $q_{t+1} = S_j$ according to A i.e. a_{ij}
  - Set $t = t + 1$
  - Repeat until $t > T$

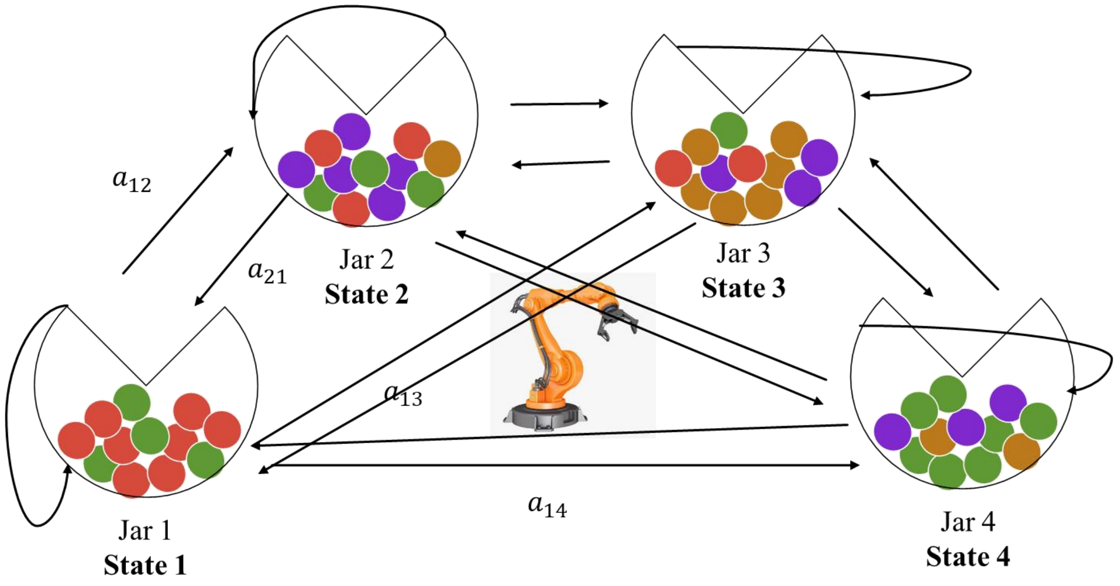
The Hidden Markov Model



- $\text{HMM} = \{N, M, A, B, \pi\}$
- $O = \{O_1, O_2, \dots, O_T\}$

- Observable Markov Model is a special case of HMM with only one non zero observation probability at each state
  - In OMM, observation distribution has only one non zero probability

What Kind of Questions Can We Answer?



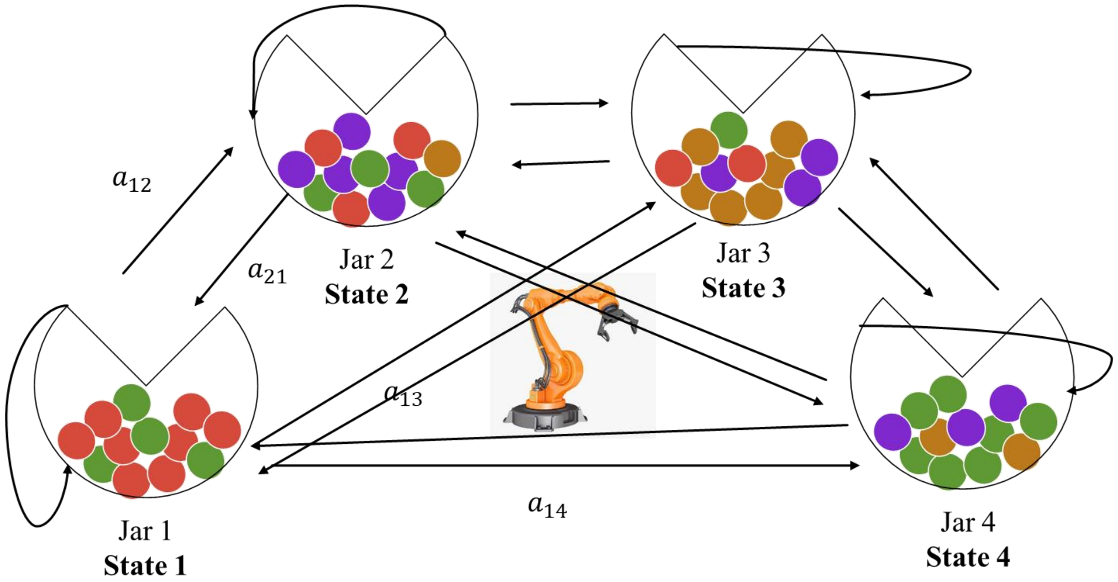
- Suppose, we have a model λ
- What is the probability that this model generates an observation sequence

$$O = \{O_1, O_2, \dots, O_T\}$$

for example, given our model, what is the probability of observing the following sequence

Given the model means given the information $\boldsymbol{\lambda = \{A, B, \pi\}}$

What Kind of Questions Can We Answer?



Given the model means given the information $\lambda = (A, B, \pi)$

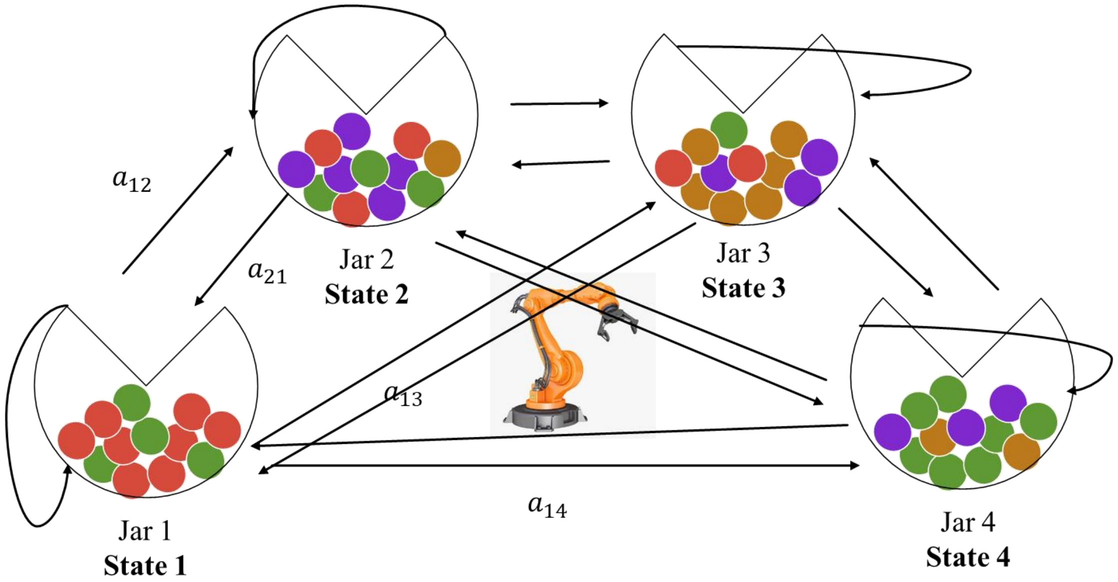
- Suppose, we have a model λ
- What is the probability that this model generates an observation sequence

$$O = \{O_1, O_2, \ldots, O_T\}$$

for example, given our model, what is the probability of observing the following sequence

What is $P(O|\lambda)$?

What Kind of Questions Can We Answer?



Given the information $\boldsymbol{\lambda = (A, B, \pi)}$

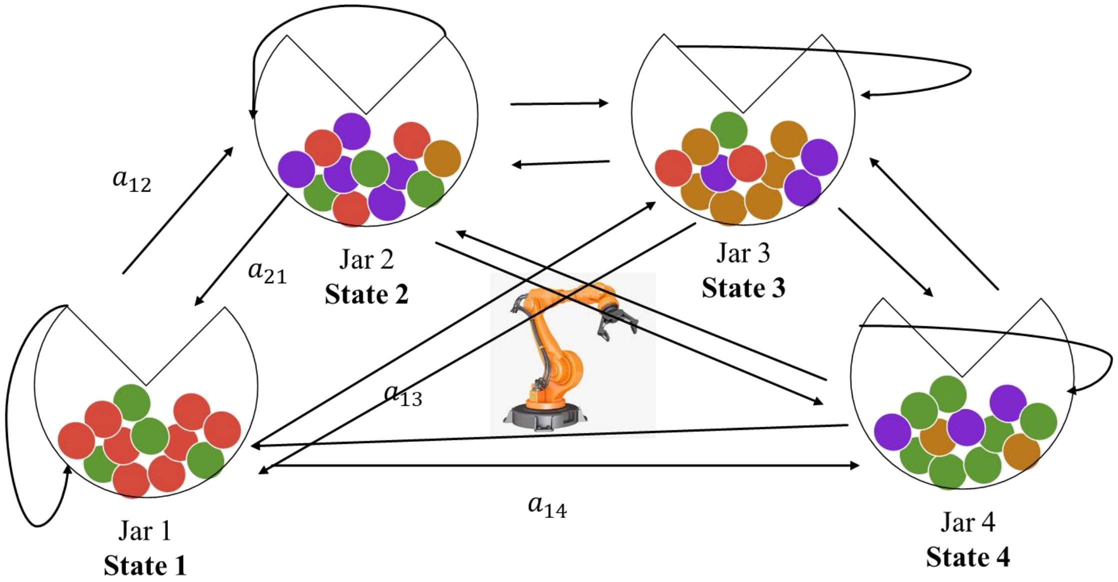
- Suppose, we have an observation sequence

  $O = \{O_1, O_2, \dots, O_T\}$

- Given the model, what sequence of states best explains the above observation?

What is $\mathbf{Q = \{q_1, q_2, \dots, q_T\}}$?

What Kind of Questions Can We Answer?



Given $\boldsymbol{O} = \{\boldsymbol{O_1}, \boldsymbol{O_2}, \ldots, \boldsymbol{O_T}\}$

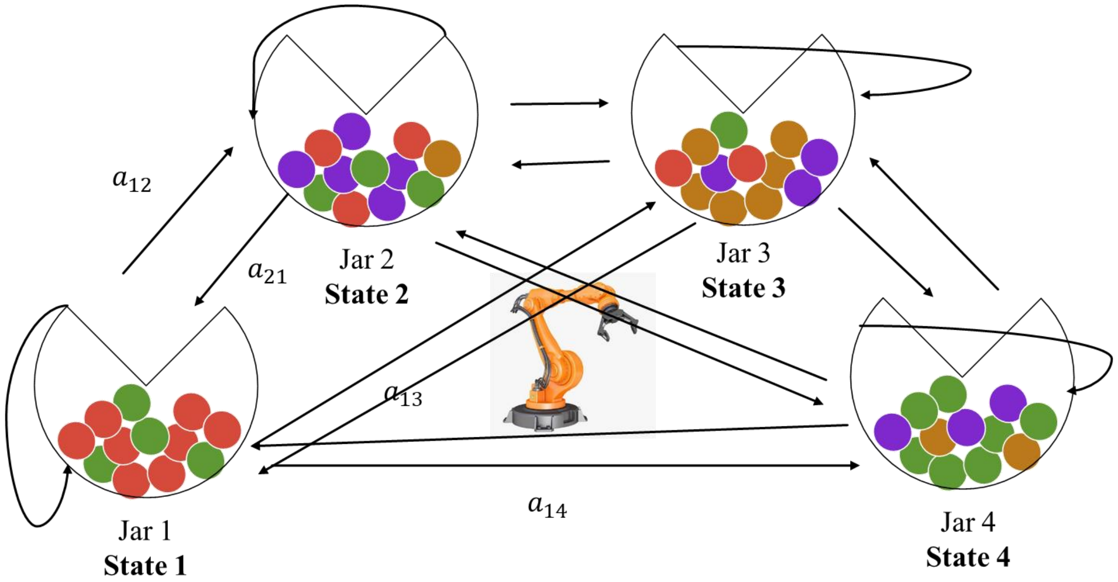
What is $\boldsymbol{\lambda} = \{\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}\}$?

- Given an observation sequence

  $O = \{O_1, O_2, \ldots, O_T\}$

- How to learn the model parameters that will maximize the chance of generating the above sequence?

Question 1



Given the model means given the information $\boldsymbol{\lambda = (A, B, \pi)}$

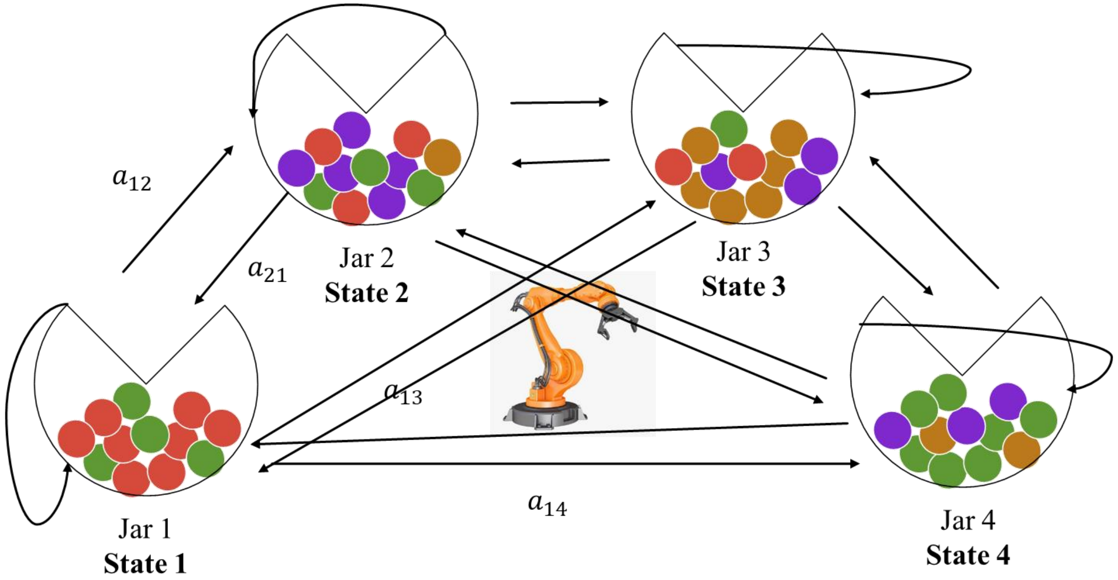
- Suppose, we have a model λ
- What is the probability that this model generates an observation sequence

$$O = \{O_1, O_2, \ldots, O_T\}$$

for example, given our model, what is the probability of observing the following sequence

What is $\boldsymbol{P(O|\lambda)}$?

Question 1

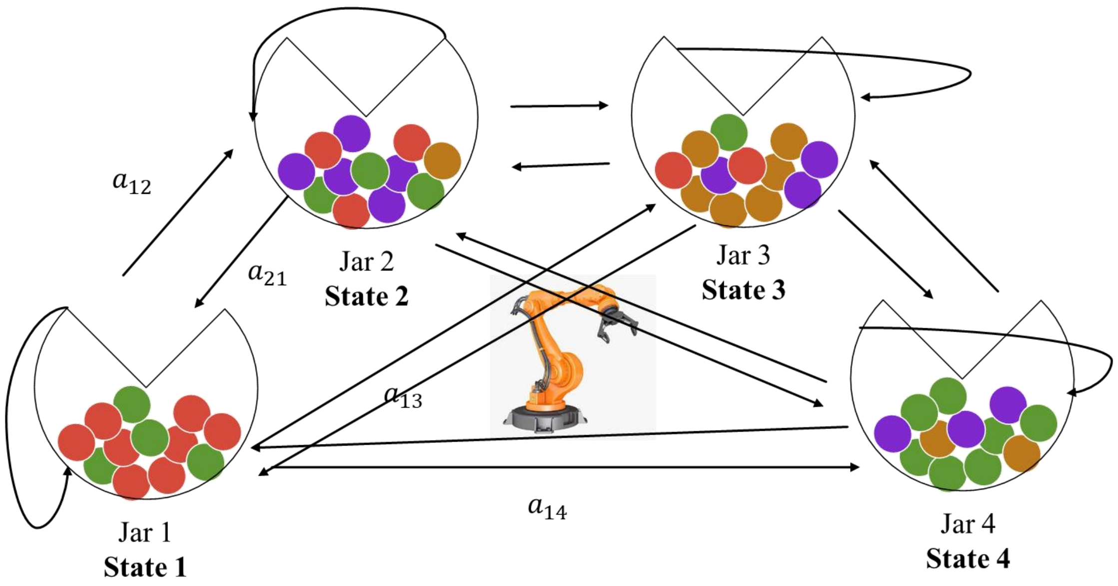


Given the model means given the information $\boldsymbol{\lambda = (A, B, \pi)}$

- Suppose, we have three models (three different sets of jars with different number of balls of these four colours, different state transition probabilities)
  - λ_1
  - λ_2
  - λ_3

Which model is more likely to generate the above observation?

Question 1



Given the model means given the information $\boldsymbol{\lambda = (A, B, \pi)}$

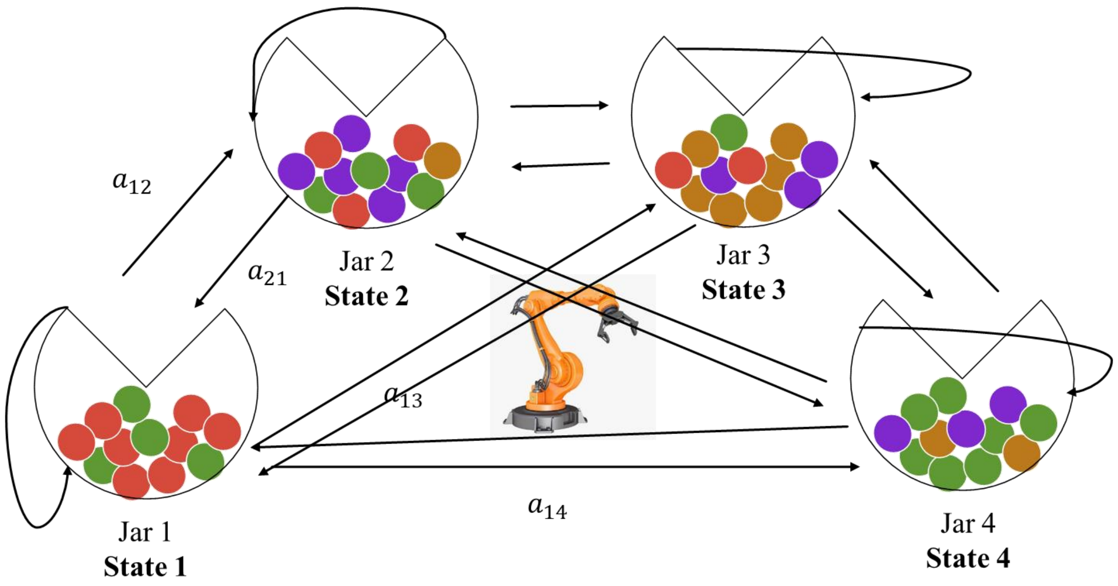
We may calculate $\boldsymbol{P(O|\lambda_1), P(O|\lambda_2), P(O|\lambda_3)}$

Find out which of these probabilities is maximum and decide which model is most likely to generate the observed sequence

- Suppose, we have three models (three different sets of jars with different number of balls of these four colours, different state transition probabilities)
  - λ_1
  - λ_2
  - λ_3

Which model is more likely to generate the above observation?

Question 1



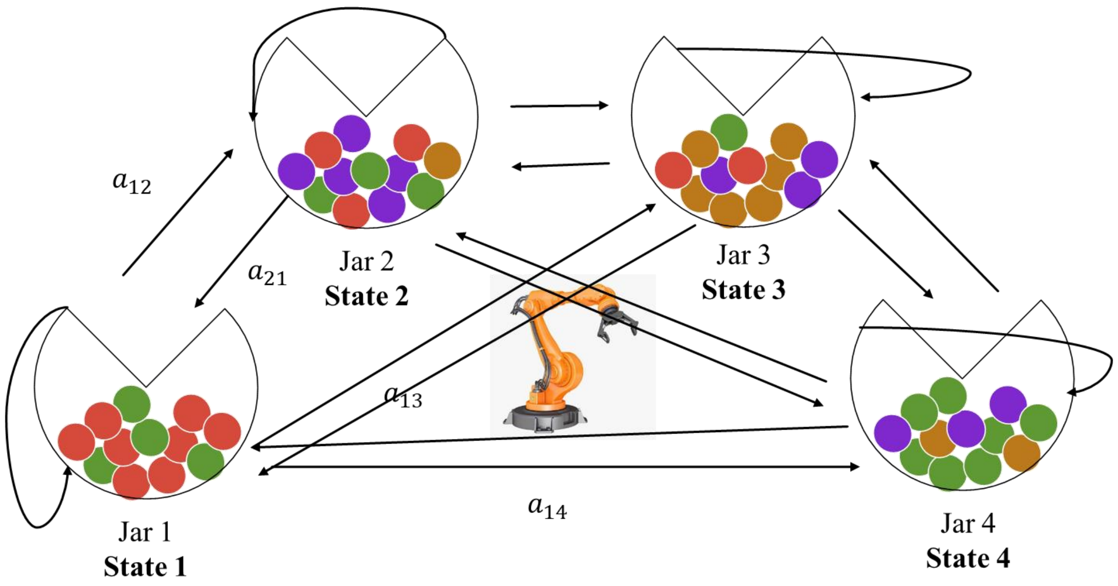
- Let's consider all possible state sequence

$$Q = q_1, q_2, \dots, q_T$$

- $O = \{O_1, O_2, \dots, O_T\}, \lambda = (A, B, \pi)$
- Probability of above observation given the state sequence

$$P(O|Q,\lambda) = P(O_1|q_1,\lambda)P(O_2|q_2,\lambda) \dots P(O_T|q_T,\lambda)$$

Question 1



- Let's consider a possible state sequence

$$Q_i = q_{1i}, q_{2i}, \dots, q_{Ti}$$

- $O = \{O_1, O_2, \dots, O_T\}, \lambda = (A, B, \pi)$
- Probability of above observation given the state sequence

$$P(O|Q_i, \lambda) = P(O_1|q_{1i}, \lambda)P(O_2|q_{2i}, \lambda) \dots P(O_T|q_{Ti}, \lambda) = \prod_{t=1}^{T} P(O_t|q_{ti}, \lambda)$$

Question 1

- Let's consider a possible state sequence

$$Q_i = q_{1i}, q_{2i}, \dots, q_{Ti}$$

- $O = \{O_1, O_2, \dots, O_T\}, \lambda = (A, B, \pi)$
- Probability of above observation given the state sequence

$$P(O|Q_i, \lambda) = P(O_1|q_{1i}, \lambda)P(O_2|q_{2i}, \lambda) \dots P(O_T|q_{Ti}, \lambda) = \prod_{t=1}^{T} P(O_t|q_{ti}, \lambda)$$

- $P(O_1|q_{1i}, \lambda) = b_{q_{1i}}(O_1), \quad P(O_2|q_{2i}, \lambda) = b_{q_{2i}}(O_2), \quad P(O_T|q_{Ti}, \lambda) = b_{q_{Ti}}(O_T)$

Question 1

- Let's consider a possible state sequence

$$Q_i = q_{1i}, q_{2i}, \ldots, q_{Ti}$$

- $O = \{O_1, O_2, \ldots, O_T\}, \lambda = (A, B, \pi)$
- Probability of above observation given the state sequence

$$P(O|Q_i, \lambda) = P(O_1|q_{1i}, \lambda)P(O_2|q_{2i}, \lambda) \ldots P(O_T|q_{Ti}, \lambda) = \prod_{t=1}^{T} P(O_t|q_{ti}, \lambda)$$

- $P(O_1|q_{1i}, \lambda) = b_{q_{1i}}(O_1), \quad P(O_2|q_{2i}, \lambda) = b_{q_{2i}}(O_2), \quad P(O_T|q_{Ti}, \lambda) = b_{q_{Ti}}(O_T)$

- $P(O|Q_i, \lambda) = b_{q_{1i}}(O_1) b_{q_{2i}}(O_2) \ldots b_{q_{Ti}}(O_T)$

Question 1

- Let's consider a possible state sequence

$$Q_i = q_{1i}, q_{2i}, \dots, q_{Ti}$$

- $O = \{O_1, O_2, \dots, O_T\}, \lambda = (A, B, \pi)$
- Probability of above observation given the state sequence

$$P(O|Q_i, \lambda) = P(O_1|q_{1i}, \lambda)P(O_2|q_{2i}, \lambda) \dots P(O_T|q_{Ti}, \lambda) = \prod_{t=1}^{T} P(O_t|q_{ti}, \lambda)$$

- $P(O|Q_i, \lambda) = b_{q_{1i}}(O_1) b_{q_{2i}}(O_2) \dots b_{q_{Ti}}(O_T)$

- Probability of observing the above state transition given the model λ

$$P(Q_i|\lambda) = \pi_{q_{1i}} a_{q_{1i}q_{2i}} a_{q_{2i}q_{3i}} \dots a_{q_{(T-1)i}q_{Ti}}$$

Question 1

- Probability of seeing the observation and the specific state transitions is
$$P(O|Q_i,\lambda)P(Q_i|\lambda) = P(O,Q_i|\lambda)$$

- Probability of seeing the observation considering specific state transitions Q_1 is
$$P(O,Q_1|\lambda) = P(O|Q_1\ \lambda)P(Q_1|\lambda)$$
- Probability of seeing the observation considering specific state transitions Q_2 is
$$P(O,Q_2|\lambda) = P(O|Q_2\ \lambda)P(Q_2|\lambda)$$

…

…

Question 1

- Probability of seeing the observation and the specific state transitions is

$$P(O|Q_i, \lambda)P(Q_i|\lambda) = P(O, Q_i|\lambda)$$

- Probability of seeing the observation considering specific state transitions Q_1 is

$$P(O, Q_1|\lambda) = P(O|Q_1, \lambda)P(Q_1|\lambda)$$

- Probability of seeing the observation considering specific state transitions Q_2 is

$$P(O, Q_2|\lambda) = P(O|Q_2, \lambda)P(Q_2|\lambda)$$

…

…

- Probability of seeing the observation regardless of any specific state transitions is

$$P(O\ |\lambda) = \sum_{\forall i} P(O|Q_i, \lambda)P(Q_i|\lambda)$$

Law of total probability

Question 1

- $P(O|Q_i, \lambda) = b_{q_{1i}}(O_1) b_{q_{2i}}(O_2) \dots b_{q_{Ti}}(O_T)$
- Probability of observing the above state transition given the model λ

$$P(Q_i|\lambda) = \pi_{q_{1i}} a_{q_{1i}q_{2i}} a_{q_{2i}q_{3i}} \dots a_{q_{(T-1)i}q_{Ti}}$$

- Probability of seeing the observation regardless of any specific state transitions is

$$P(O\,|\lambda) = \sum_{\forall i} P(O|Q_i, \lambda) P(Q_i|\lambda)$$

$$= \sum_{\forall i} \pi_{q_{1i}} b_{q_{1i}}(O_1) a_{q_{1i}q_{2i}} b_{q_{2i}}(O_2) a_{q_{2i}q_{3i}} \dots a_{q_{(T-1)i}q_{Ti}} b_{q_{Ti}}(O_T)$$

Calculation

- Probability of seeing the observation regardless of any specific state transitions is

$$= \sum_{\forall i} \pi_{q_{1i}} b_{q_{1i}}(O_1) a_{q_{1i}q_{2i}} b_{q_{2i}}(O_2) a_{q_{2i}q_{3i}} \dots a_{q_{(T-1)i}q_{Ti}} b_{q_{Ti}}(O_T)$$

- How many possible state sequence: N^T
- How many multiplications per state sequence: $(2T-1)$
- Total number of operations: $(2T-1)N^T + (N^T - 1)$

Calculation

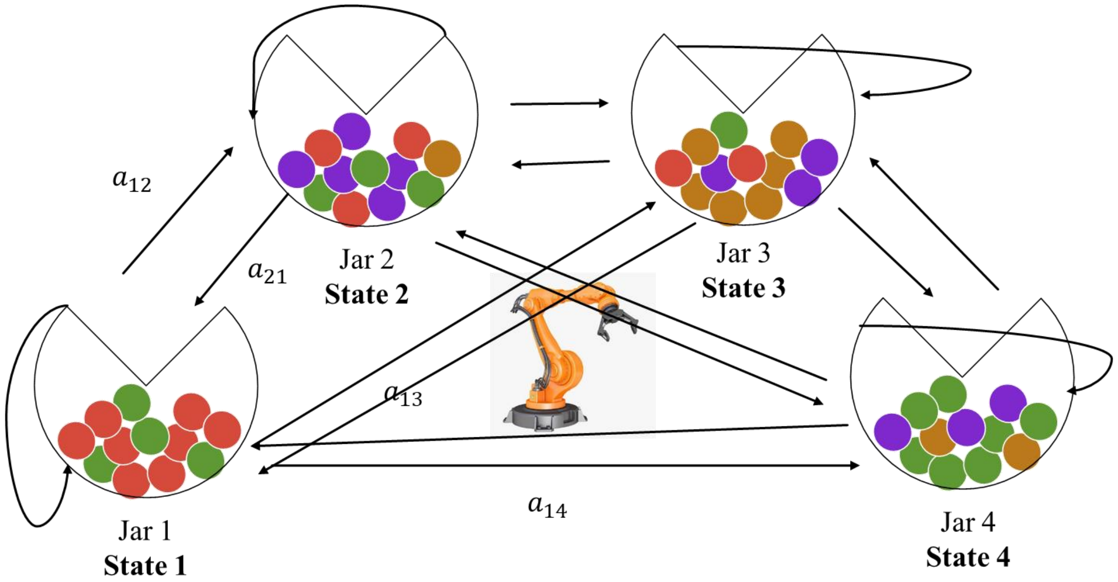
- Probability of seeing the observation regardless of any specific state transitions is

$$= \sum_{\forall i} \pi_{q_{1i}} b_{q_{1i}}(O_1) a_{q_{1i}q_{2i}} b_{q_{2i}}(O_2) a_{q_{2i}q_{3i}} \dots a_{q_{(T-1)i}q_{Ti}} b_{q_{Ti}}(O_T)$$

- How many possible state sequence: N^T
- How many multiplications per state sequence: $(2T - 1)$
- Total number of operations: $(2T - 1)N^T + (N^T - 1)$

- If total number of observations $T = 100$, number of states $N = 10$
  - Total number of operations $\sim 10^{100}$

Question 1



Given the model means given the information $\boldsymbol{\lambda = (A, B, \pi)}$

- Suppose, we have a model λ
- What is the probability that this model generates an observation sequence

$$O = \{O_1, O_2, \dots, O_T\}$$

for example, given our model, what is the probability of observing the following sequence

What is $\boldsymbol{P(O|\lambda)}$?

Calculation

- Probability of seeing the observation regardless of any specific state transitions is

$$= \sum_{\forall i} \pi_{q_{1i}} b_{q_{1i}}(O_1) a_{q_{1i}q_{2i}} b_{q_{2i}}(O_2) a_{q_{2i}q_{3i}} \dots a_{q_{(T-1)i}q_{Ti}} b_{q_{Ti}}(O_T)$$

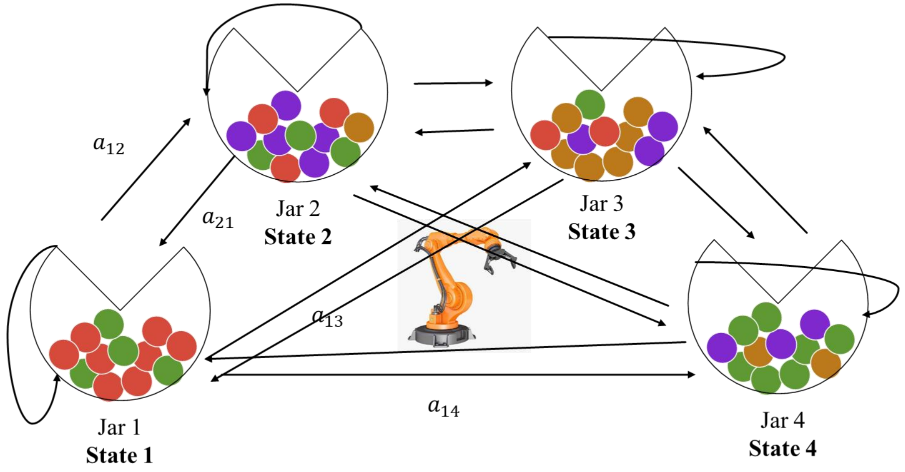
- How many possible state sequence: N^T
- How many multiplications per state sequence: $(2T - 1)$
- Total number of operations: $(2T - 1)N^T + (N^T - 1)$

- If total number of observations $T = 100$, number of states $N = 10$
  - Total number of operations $\sim 10^{100}$

Solution: Forward backward Algorithm

- Probability of seeing the observation regardless of any specific state transitions is

$$= \sum_{\forall i} \pi_{q_{1i}} b_{q_{1i}}(O_1) a_{q_{1i}q_{2i}} b_{q_{2i}}(O_2) a_{q_{2i}q_{3i}} \dots a_{q_{(T-1)i}q_{Ti}} b_{q_{Ti}}(O_T)$$



- Helper function (to reduce the number of repeating calculations)

$$\alpha_t(i) = P(O_1, O_2, \dots O_t, q_t = S_i | \lambda)$$

$\alpha_t(i)$ is the probability of seeing observations $O_1, O_2, \dots O_t$ and reaching at state S_i at t given our model

Solution: Forward backward Algorithm

- Helper function (to reduce the number of repeating calculations)

$$\alpha_t(i) = P(O_1, O_2, \ldots O_t, q_t = S_i|\lambda)$$

$\alpha_t(i)$ is the probability of seeing observations $O_1, O_2, \ldots O_t$ and reaching at state S_i at t given our model

- Inductive solution
  - Base case: $\alpha_1(i) = P(O_1, q_1 = S_i|\lambda) = \pi_i b_i(O_1) \qquad 1 \leq i \leq N$

(probability of starting at state i and seeing O_1)

Solution: Forward backward Algorithm

- Helper function (to reduce the number of repeating calculations)

$$\alpha_t(i) = P(O_1, O_2, \dots O_t, q_t = S_i | \lambda)$$

- Inductive solution
  - Base case: $\alpha_1(i) = P(O_1, q_1 = S_i|\lambda) = \pi_i b_i(O_1)$ $\quad 1 \leq i \leq N$
  - Inductive step

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i)\, a_{ij}\right] b_j(O_{t+1}) \qquad 1 \leq i \leq N$$

$$1 \leq t \leq T - 1$$

Solution: Forward backward Algorithm

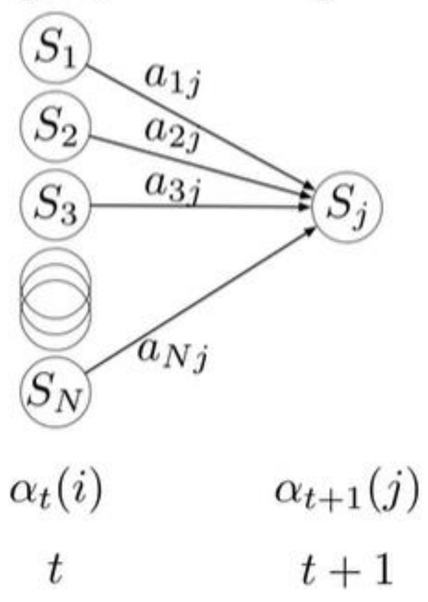
- Helper function (to reduce the number of repeating calculations)

$$\alpha_t(i) = P(O_1, O_2, \dots O_t, q_t = S_i|\lambda)$$

- Inductive solution
  - Base case: $\alpha_1(i) = P(O_1, q_1 = S_i|\lambda) = \pi_i b_i(O_1)$ $\quad 1 \le i \le N$
  - Inductive step

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i)\, a_{ij}\right] b_j(O_{t+1}) \qquad 1 \le i \le N$$

$$1 \le t \le T - 1$$



Solution: Forward backward Algorithm

- Helper function (to reduce the number of repeating calculations)

$$\alpha_t(i) = P(O_1, O_2, \ldots O_t, q_t = S_i|\lambda)$$

- Inductive solution
  - Base case: $\alpha_1(i) = P(O_1, q_1 = S_i|\lambda) = \pi_i b_i(O_1)$ $\qquad 1 \leq i \leq N$
  - Inductive step

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i)\, a_{ij}\right] b_j(O_{t+1}) \qquad 1 \leq i \leq N$$

$$1 \leq t \leq T-1$$

  - Find $\alpha_T(i)$ using above equations

Solution: Forward backward Algorithm

- Helper function (to reduce the number of repeating calculations)

$$\alpha_t(i) = P(O_1, O_2, \dots O_t, q_t = S_i|\lambda)$$

- Inductive solution
  - Base case: $\alpha_1(i) = P(O_1, q_1 = S_i|\lambda) = \pi_i b_i(O_1)$ $\qquad 1 \le i \le N$
  - Inductive step

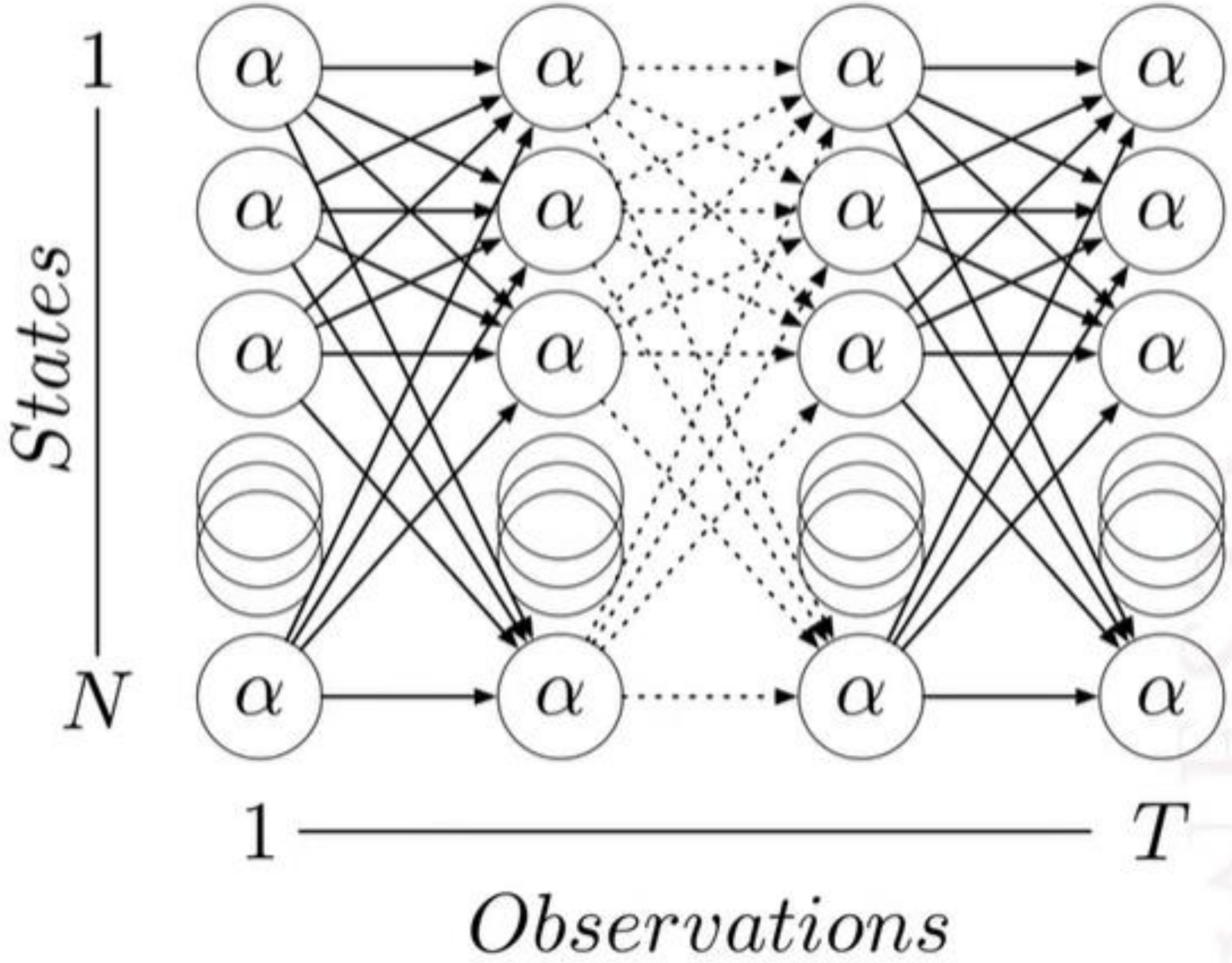
$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i)\, a_{ij}\right] b_j(O_{t+1}) \qquad 1 \le i \le N$$

$$1 \le t \le T-1$$

  - Find $\alpha_T(i)$ using above equations
  - Final step $P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$

Solution: Forward backward Algorithm

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



Solution: Forward backward Algorithm

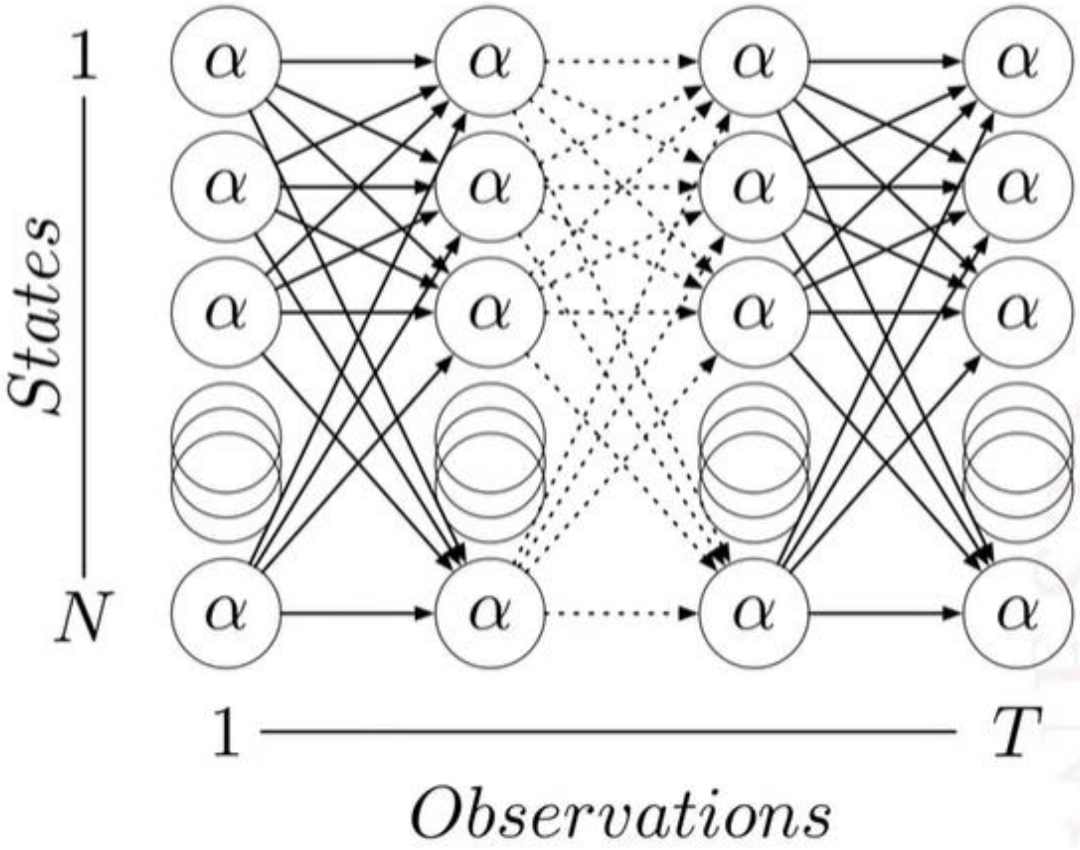
Final step $P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$

Number of computations $\sim N^2 T$

For $N = 10, T = 100$

Number of computations $\sim 10^2 \times 100 = 10^4$

Earlier, it was $\sim 10^{100}$



Forward backward Algorithm

- Helper function 1

$$\alpha_t(i) = P(O_1, O_2, \dots O_t, q_t = S_i | \lambda)$$

Forward backward Algorithm

- Helper function 1 (forward)

$$\alpha_t(i) = P(O_1, O_2, \ldots O_t, q_t = S_i|\lambda)$$

- Helper function 2 (backward)

$$\beta_t(i) = P(O_{t+1}, O_{t+2}, \ldots O_T|q_t = S_i, \lambda)$$

probability of seeing observations $O_{t+1}, O_{t+2}, \ldots O_T$ in future given that we are starting at state S_i at t and given our model

Solution: Forward backward Algorithm

- Helper function (to reduce the number of repeating calculations)

$$\beta_t(i) = P(O_{t+1}, O_{t+2}, \dots O_T | q_t = S_i, \lambda)$$

- Inductive solution
  - Base case: $\beta_T(i) = 1 \qquad\qquad 1 \leq i \leq N$
  - Inductive step

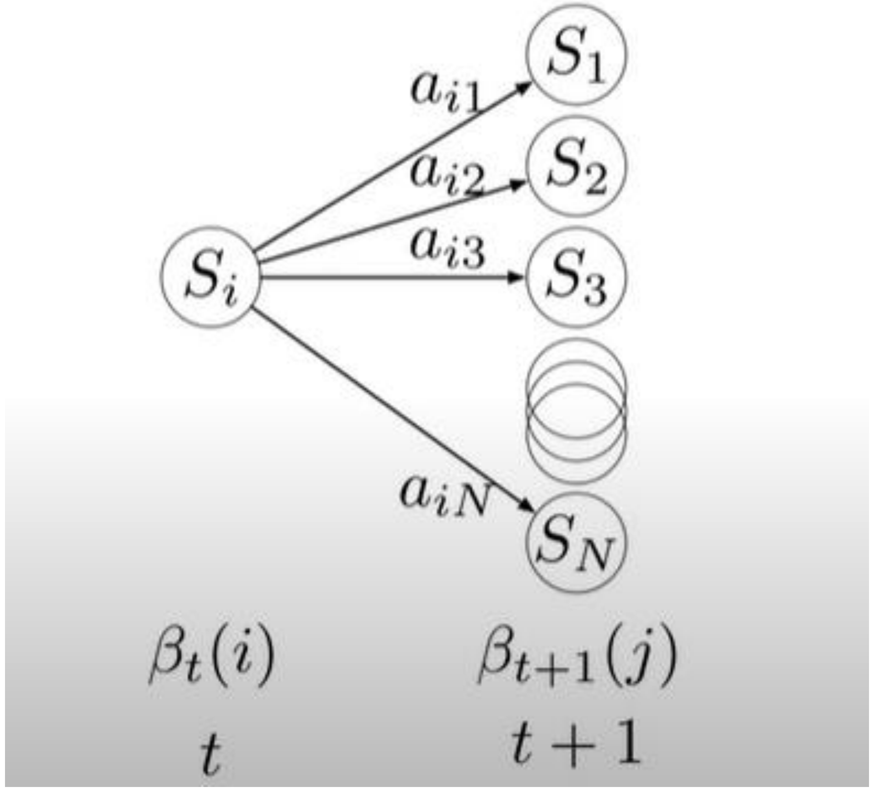
$$\beta_t(i) = \sum_{j=1}^{N} a_{ij}\, b_j(O_{t+1})\beta_{t+1}(j) \qquad 1 \leq i \leq N$$

$$t = T-1, T-2, \dots, 1$$

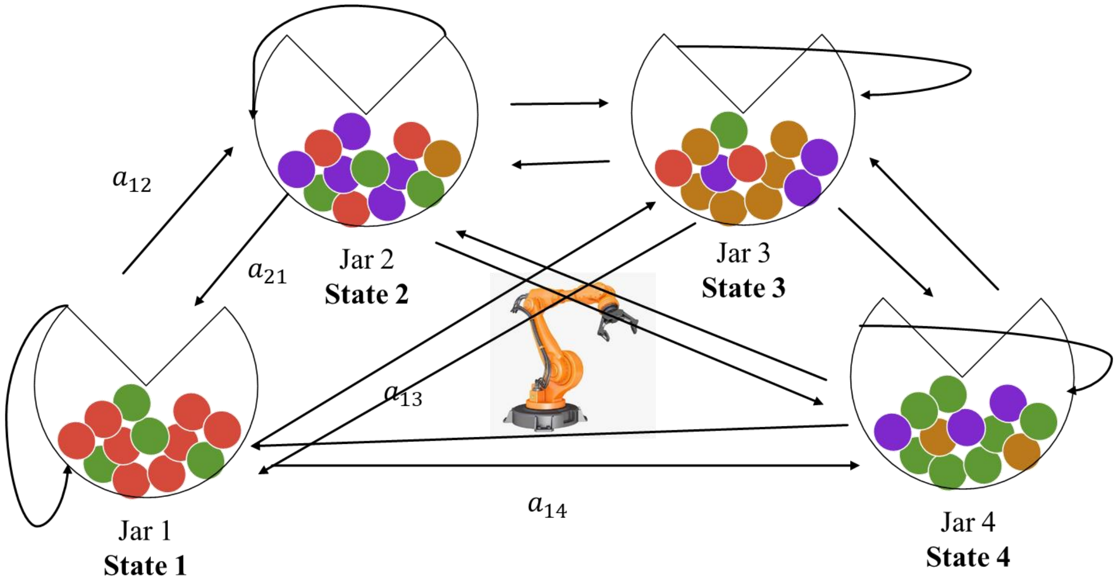
Solution: Forward backward Algorithm

- Helper function (to reduce the number of repeating calculations)

$$\beta_t(i) = P(O_{t+1}, O_{t+2}, \ldots O_T | q_t = S_i, \lambda)$$



Question 2



Given the information $\boldsymbol{\lambda = (A, B, \pi)}$

- Suppose, we have an observation sequence

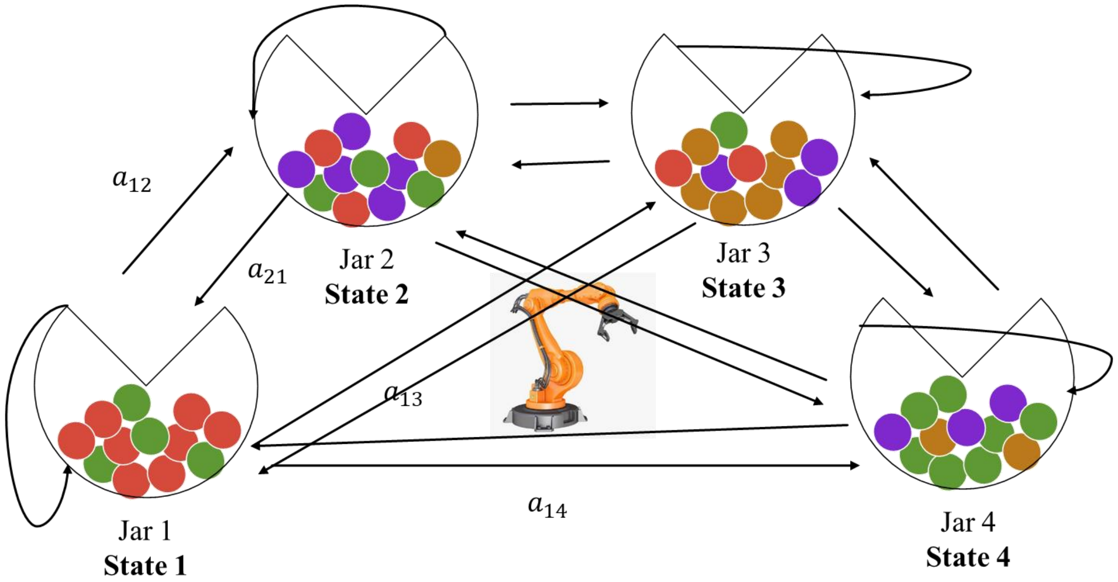
  $$O = \{O_1, O_2, \ldots, O_T\}$$

- Given the model, what sequence of states best explains the above observation?

What is $\mathbf{Q = \{q_1, q_2, \ldots, q_T\}}$?

Question 2



Given the information $\boldsymbol{\lambda = (A, B, \pi)}$

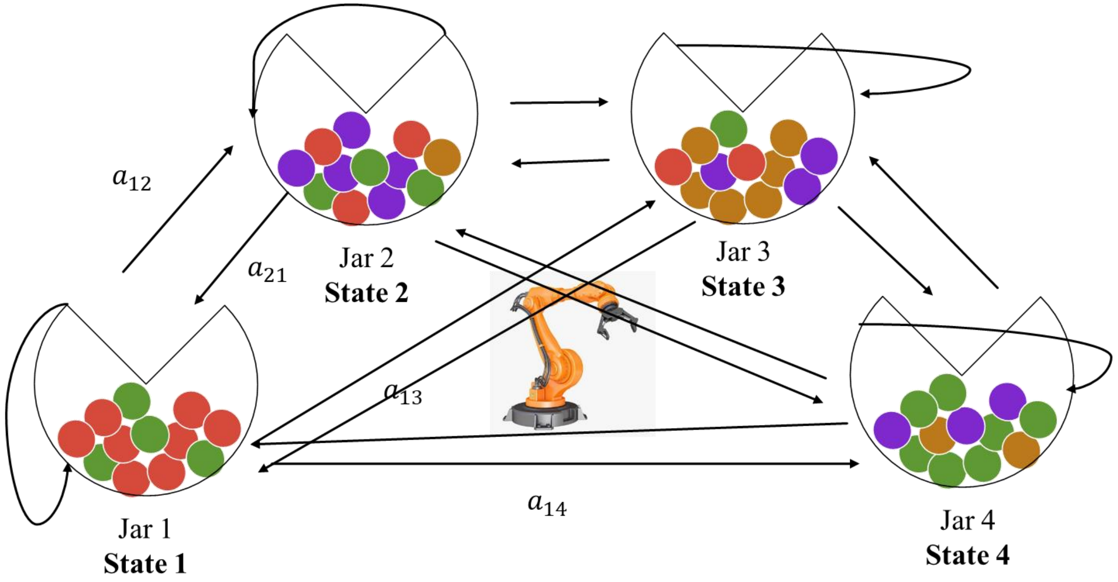
- Suppose, we have an observation sequence

  $$O = \{O_1, O_2, \dots, O_T\}$$

- Given the model, what sequence of states **best** explains the above observation?
- What is the meaning of best?

What is $\mathbf{Q = \{q_1, q_2, \dots, q_T\}}$**?**

Question 2



Given the information $\boldsymbol{\lambda = (A, B, \pi)}$

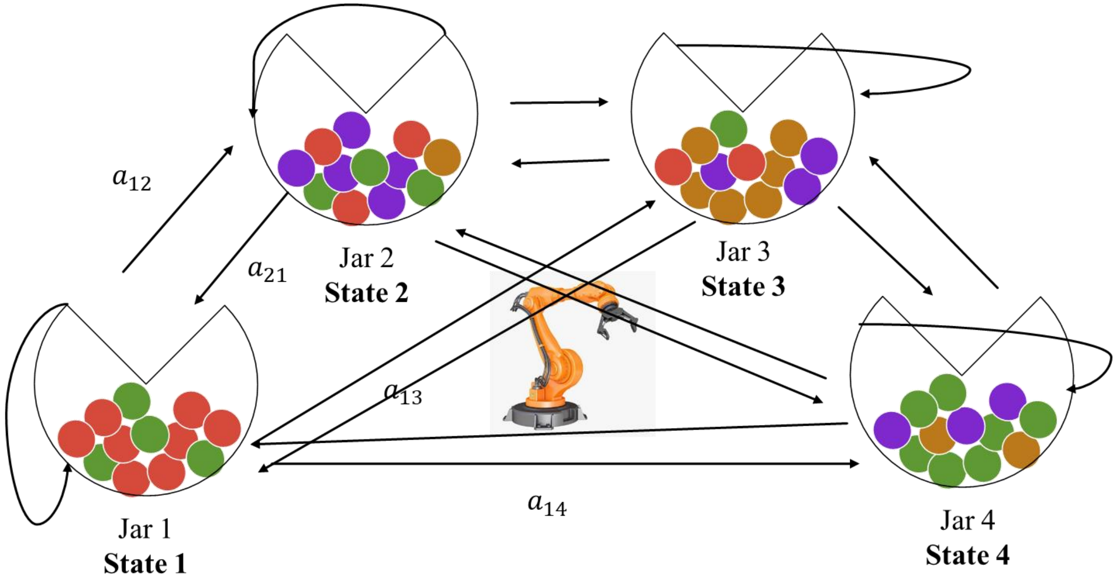
What is $\mathbf{Q = \{q_1, q_2, \dots, q_T\}}$?

- Suppose, we have an observation sequence

  $O = \{O_1, O_2, \dots, O_T\}$

- Given the model, what sequence of states **best** explains the above observation?
- What is the meaning of best?
- What is the most likely state at any given time t given the observation

Question 2



Given the information $\boldsymbol{\lambda = (A, B, \pi)}$

What is $\mathbf{Q = \{q_1, q_2, \dots, q_T\}}$?

- Suppose, we have an observation sequence

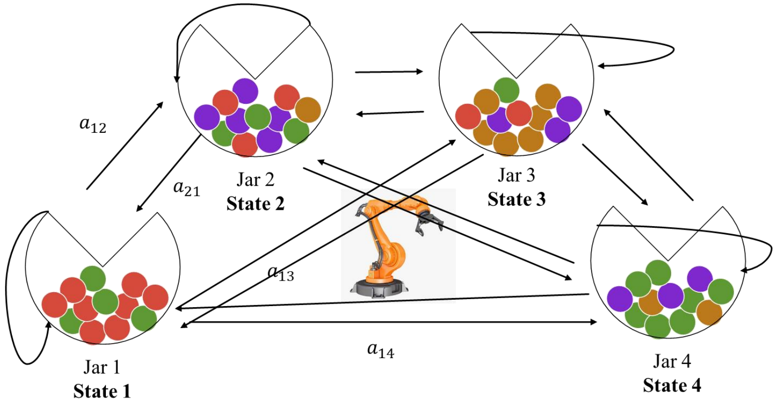
  $O = \{O_1, O_2, \dots, O_T\}$

- What is the most likely state at any given time t given the observation

- What is the probability of state i at time t given the observation

  $$\gamma_t(i) = P(q_t = S_i | O, \lambda)$$

Question 2



$$\boldsymbol{\alpha_t(i) = P(O_1, O_2, \ldots O_t, q_t = S_i|\lambda)}$$

probability of seeing observations $O_1, O_2, \ldots O_t$ and reaching at state S_i at t given our model

$$\boldsymbol{\beta_t(i) = P(O_{t+1}, O_{t+2}, \ldots O_T|q_t = S_i, \lambda)}$$

probability of seeing observations $O_{t+1}, O_{t+2}, \ldots O_T$ in future given that we are starting at state S_i at t and given our model

- Suppose, we have an observation sequence

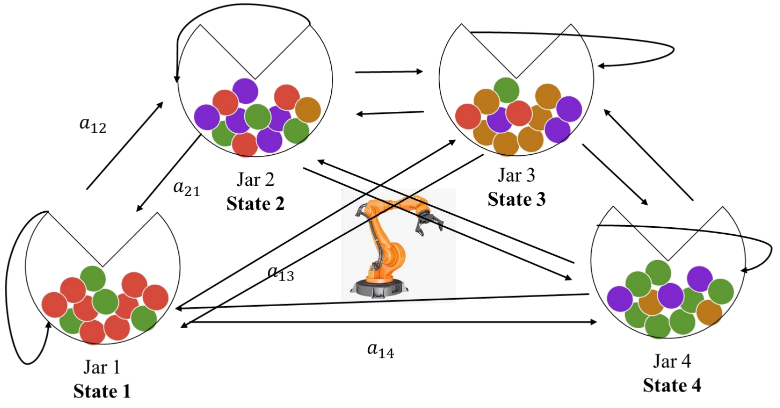
  $$O = \{O_1, O_2, \ldots, O_T\}$$

- What is the probability of state i at time t given the observation

  $$\gamma_t(i) = P(q_t = S_i|O, \lambda)$$

- $\gamma_t(i) = \dfrac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)}$

Question 2



$$\boldsymbol{\alpha_t(i) = P(O_1, O_2, \dots O_t, q_t = S_i|\lambda)}$$

probability of seeing observations $O_1, O_2, \dots O_t$ and reaching at state S_i at t given our model

$$\boldsymbol{\beta_t(i) = P(O_{t+1}, O_{t+2}, \dots O_T|q_t = S_i, \lambda)}$$

probability of seeing observations $O_{t+1}, O_{t+2}, \dots O_T$ in future given that we are starting at state S_i at t and given our model

- Suppose, we have an observation sequence

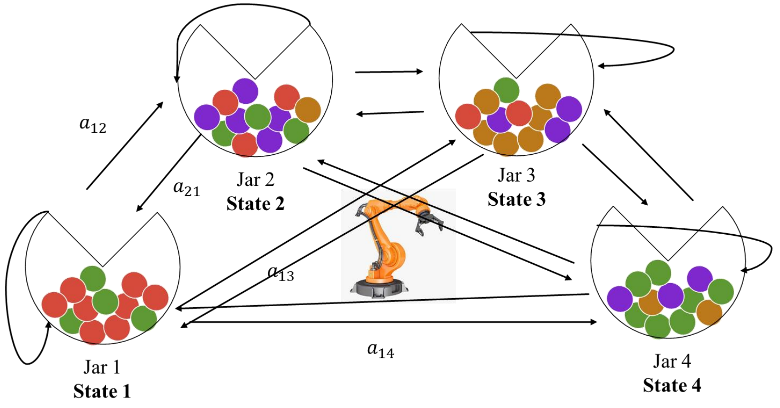
$$O = \{O_1, O_2, \dots, O_T\}$$

- What is the probability of state i at time t given the observation

$$\gamma_t(i) = P(q_t = S_i|O, \lambda)$$

- $\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)}$

Question 2



$$\boldsymbol{\alpha_t(i) = P(O_1, O_2, \ldots O_t, q_t = S_i|\lambda)}$$

probability of seeing observations $O_1, O_2, \ldots O_t$ and reaching at state S_i at t given our model

$$\boldsymbol{\beta_t(i) = P(O_{t+1}, O_{t+2}, \ldots O_T|q_t = S_i, \lambda)}$$

probability of seeing observations $O_{t+1}, O_{t+2}, \ldots O_T$ in future given that we are starting at state S_i at t and given our model

- Suppose, we have an observation sequence

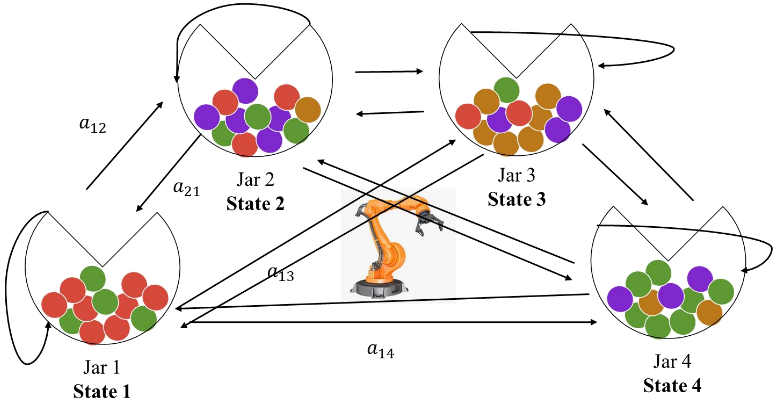
  $O = \{O_1, O_2, \ldots, O_T\}$

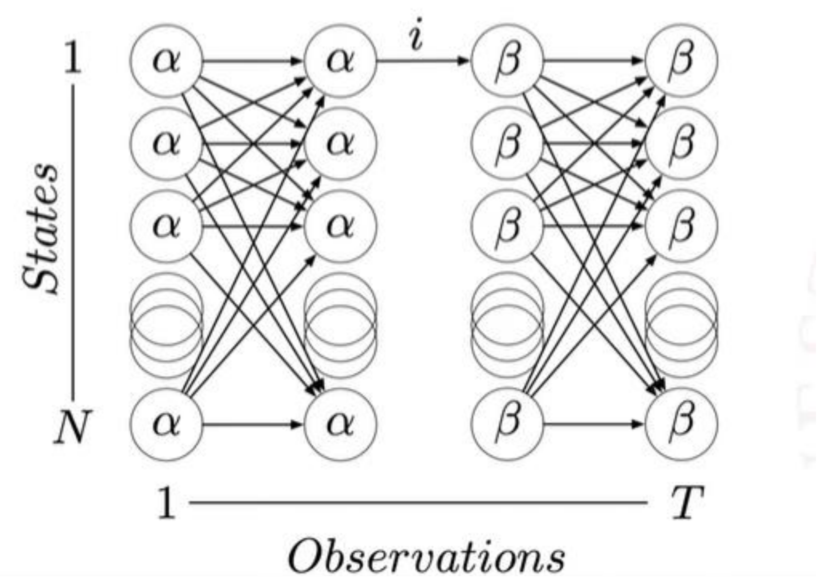
- What is the probability of state i at time t given the observation

$$\gamma_t(i) = P(q_t = S_i|O, \lambda)$$

- $\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)}$
- $\sum_{i=1}^{N} \gamma_t(i) = 1$

Question 2





- Suppose, we have an observation sequence

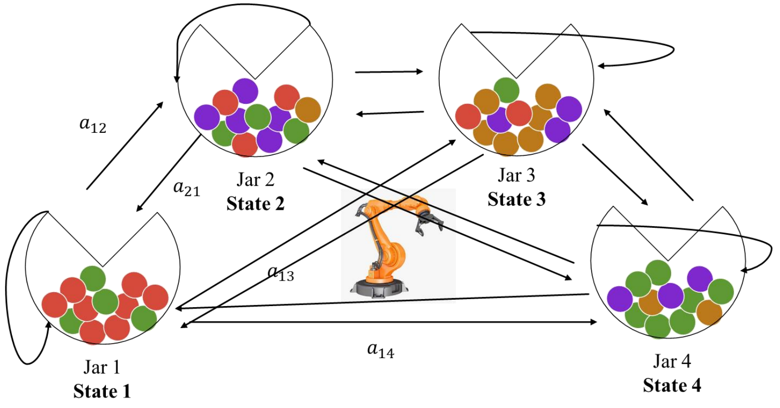
$$O = \{O_1, O_2, \dots, O_T\}$$

- What is the probability of state i at time t given the observation

$$\gamma_t(i) = P(q_t = S_i | O, \lambda)$$

- $\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)}$
- $\sum_{i=1}^{N} \gamma_t(i) = 1$

Question 2



- Suppose, we have an observation sequence

$$O = \{O_1, O_2, \ldots, O_T\}$$

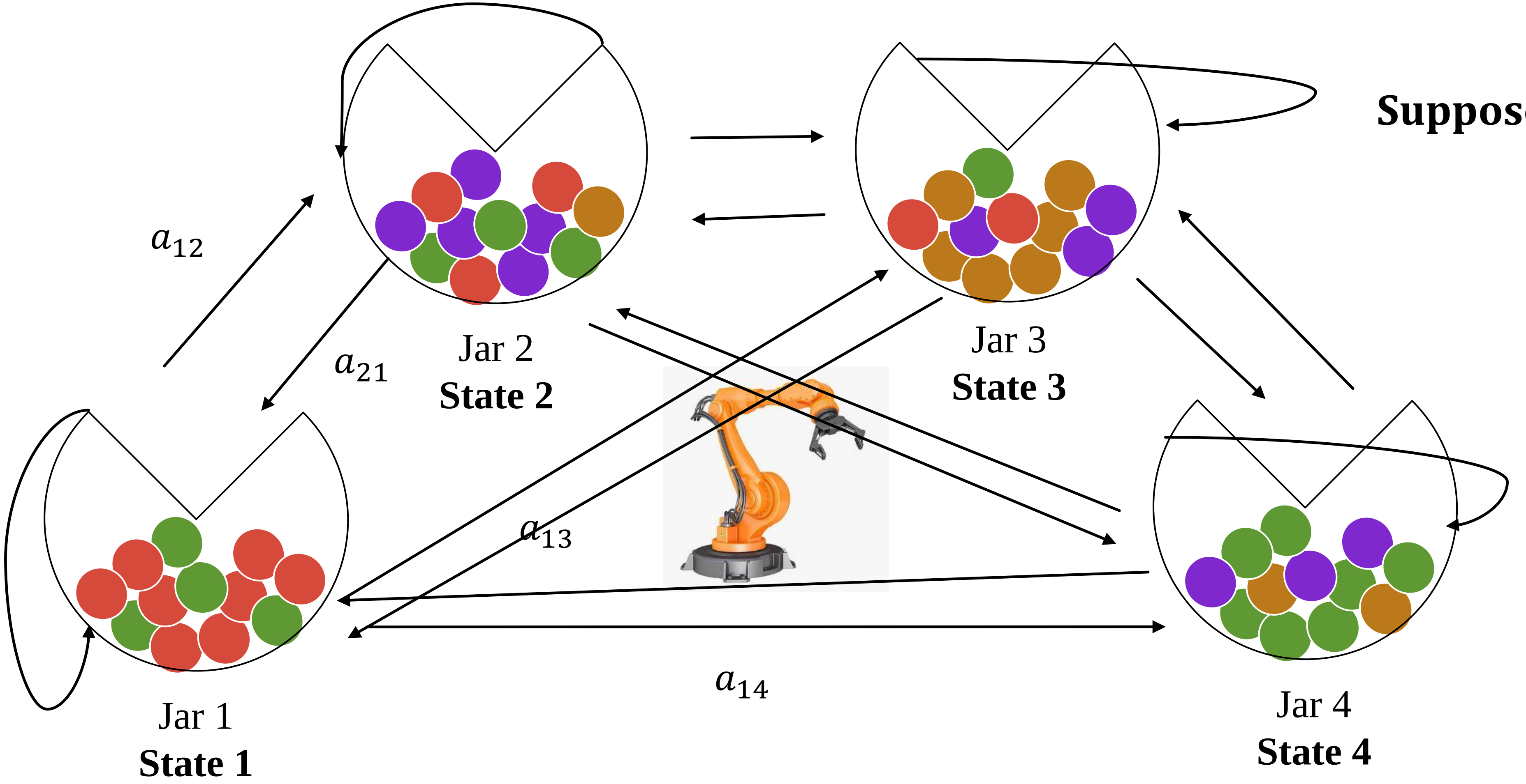
- What is the probability of state i at time t given the observation

$$\gamma_t(i) = P(q_t = S_i | O, \lambda)$$

- $\gamma_t(i) = \dfrac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} = \dfrac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)}$
- Most likely state at time t

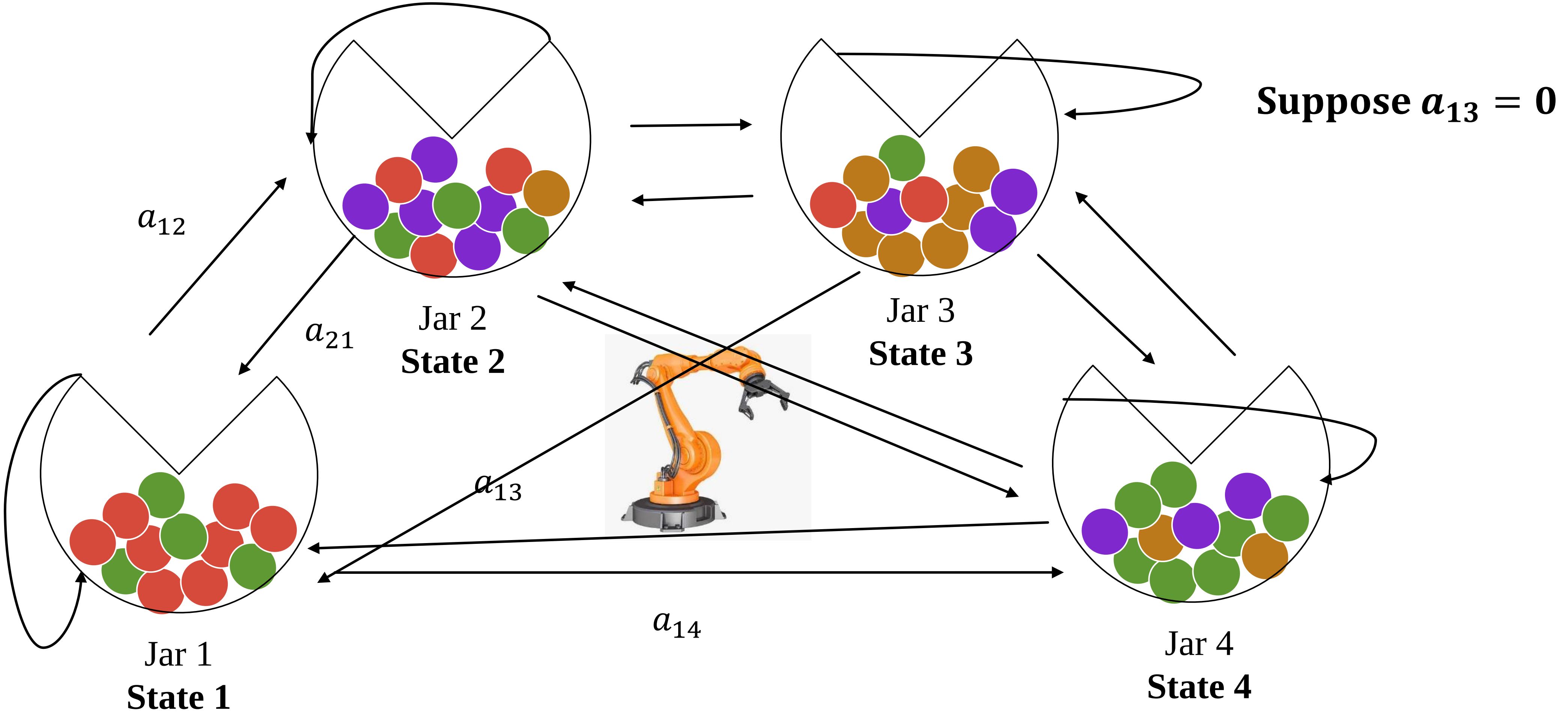
$$q_t = \underbrace{argmax}_{1 \le i \le N} [\gamma_t(i)], \qquad 1 \le t \le T$$

Problem with This Approach



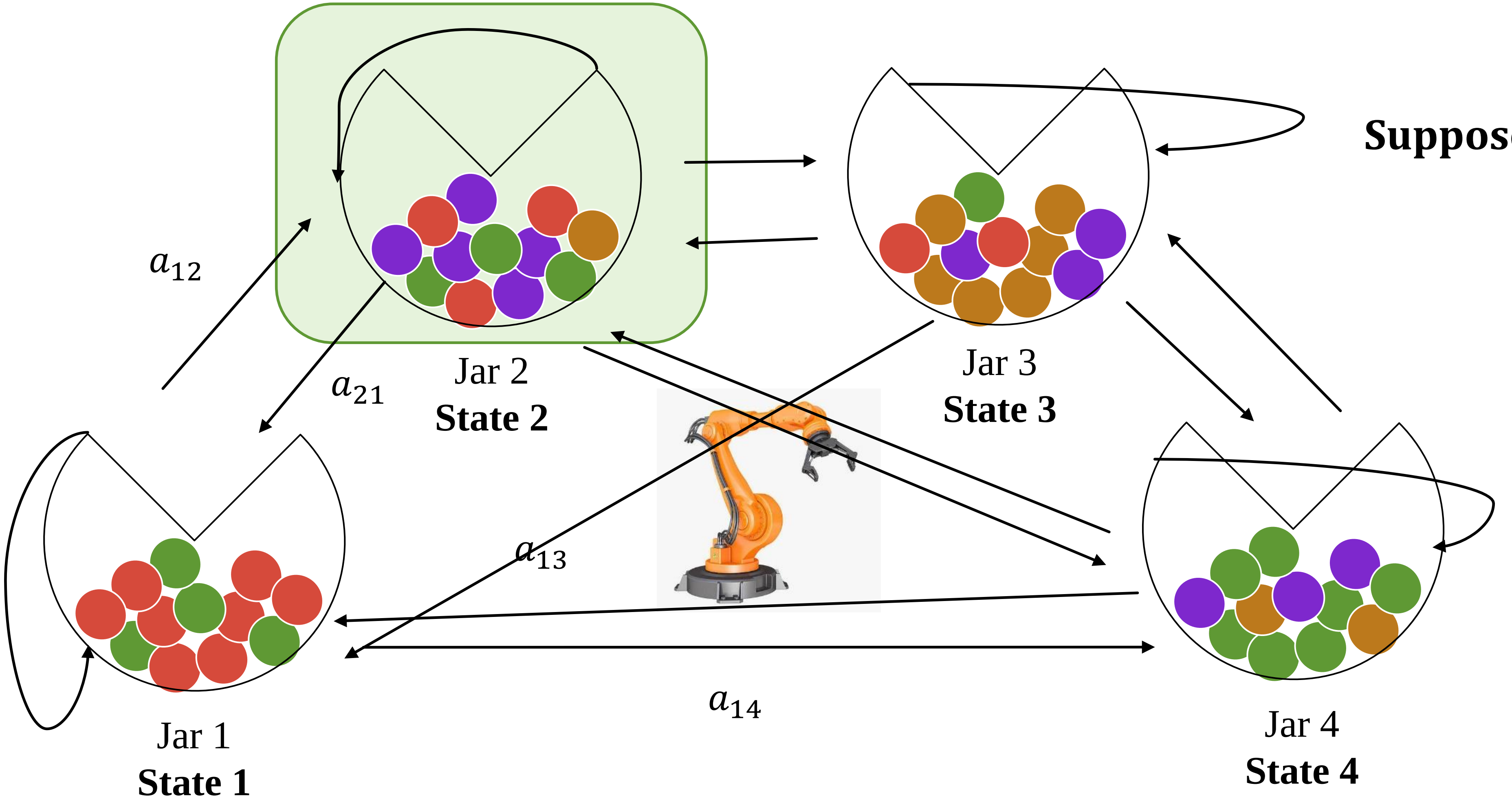
Suppose $a_{13} = 0$

Problem with This Approach



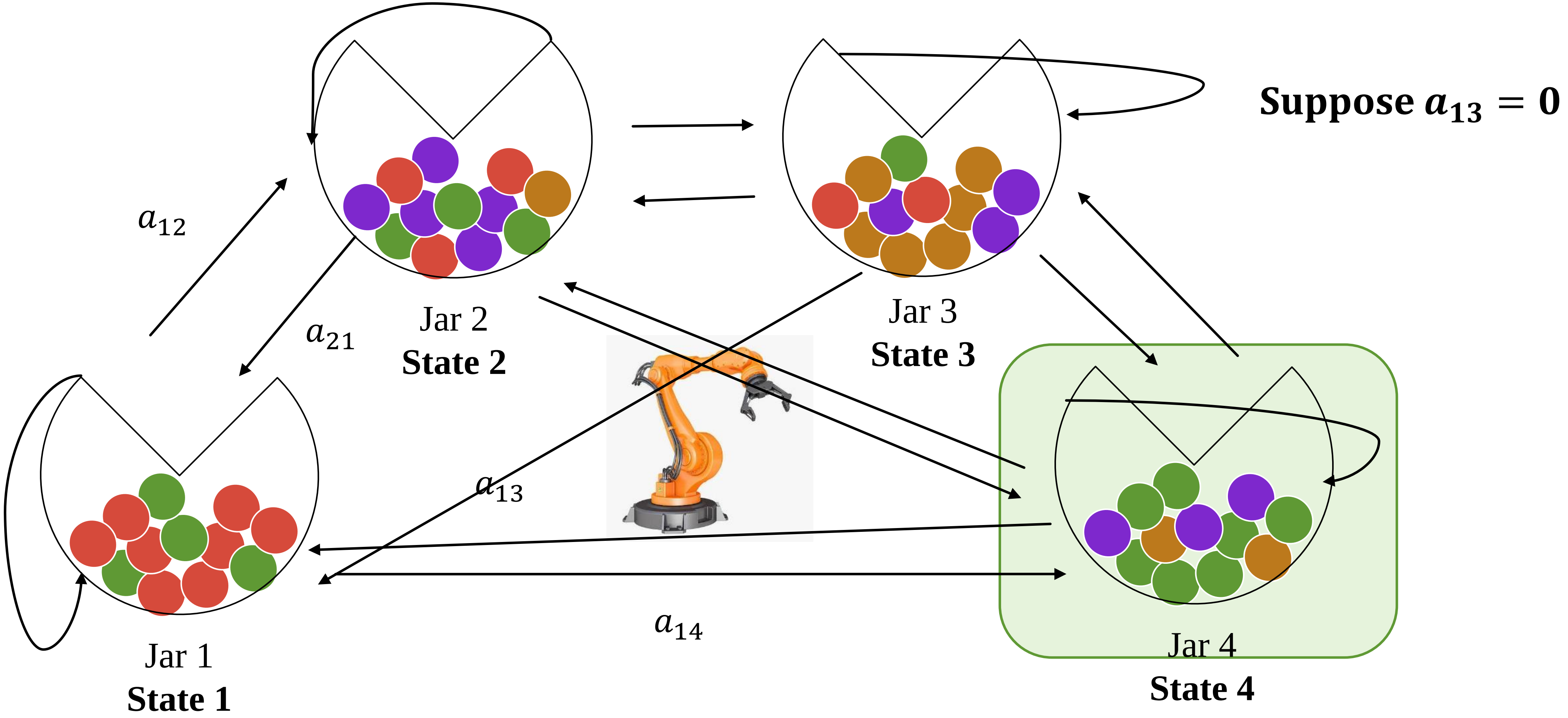
Problem with This Approach

Suppose $a_{13} = 0$



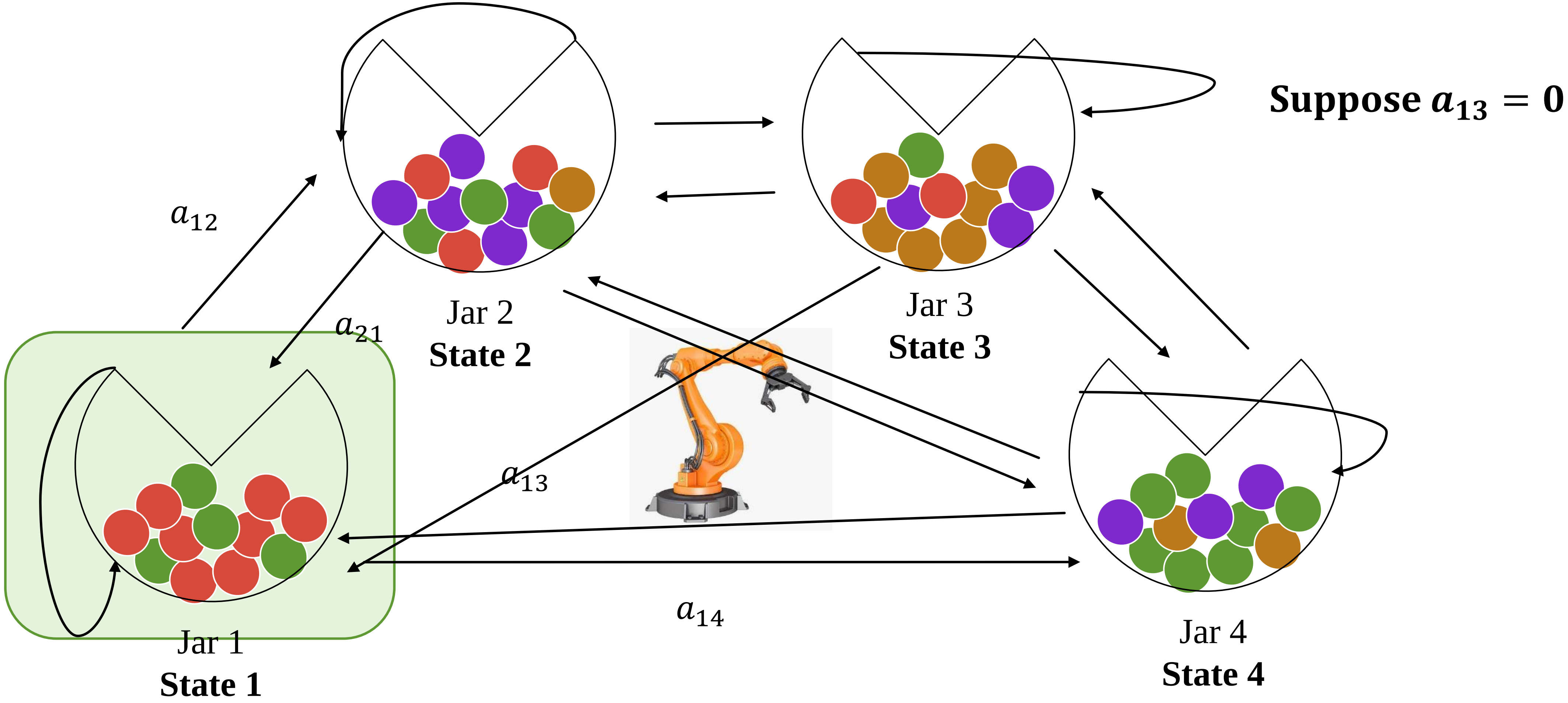
S2,

Problem with This Approach



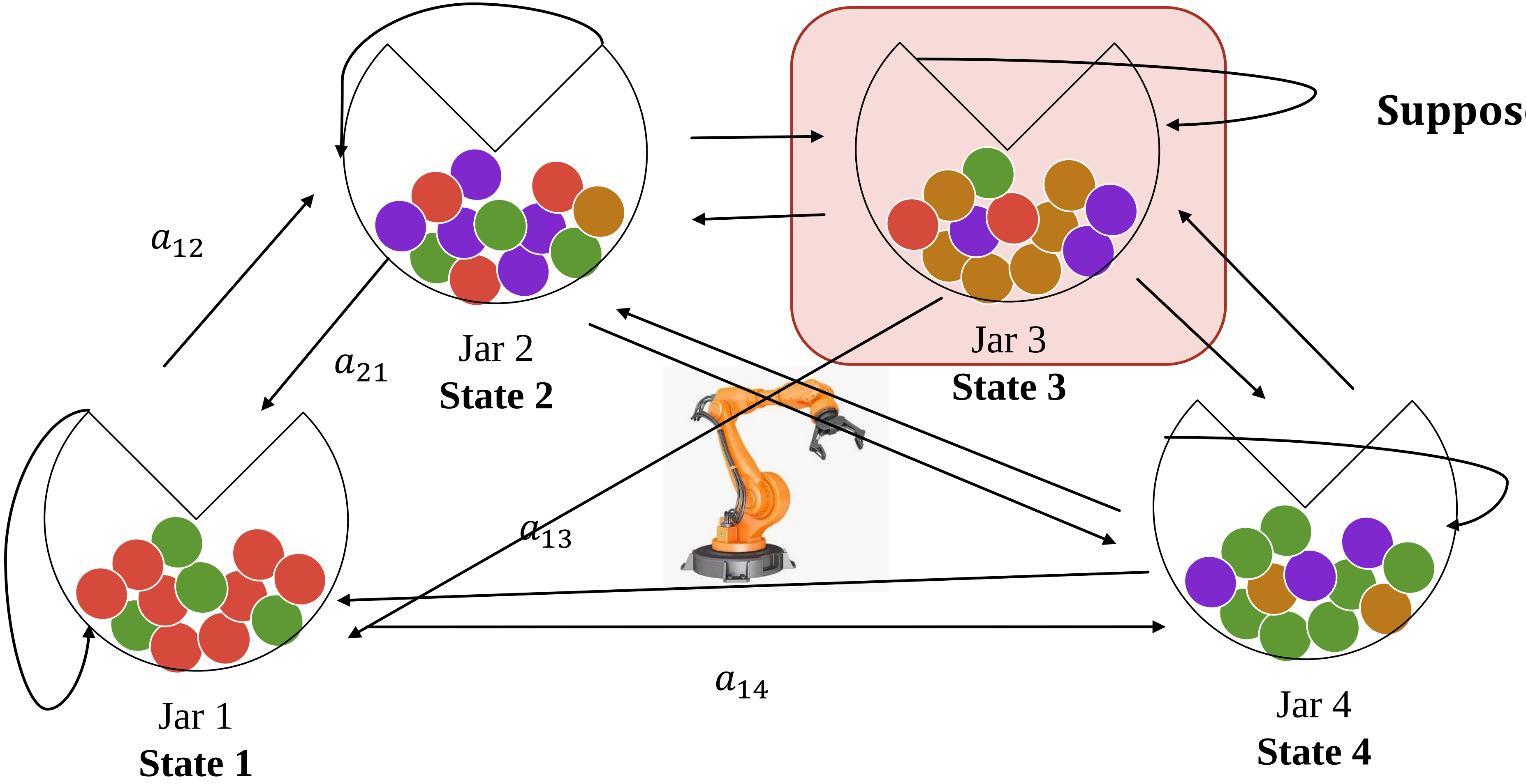
S2, S4

Problem with This Approach



S2, S4, S1

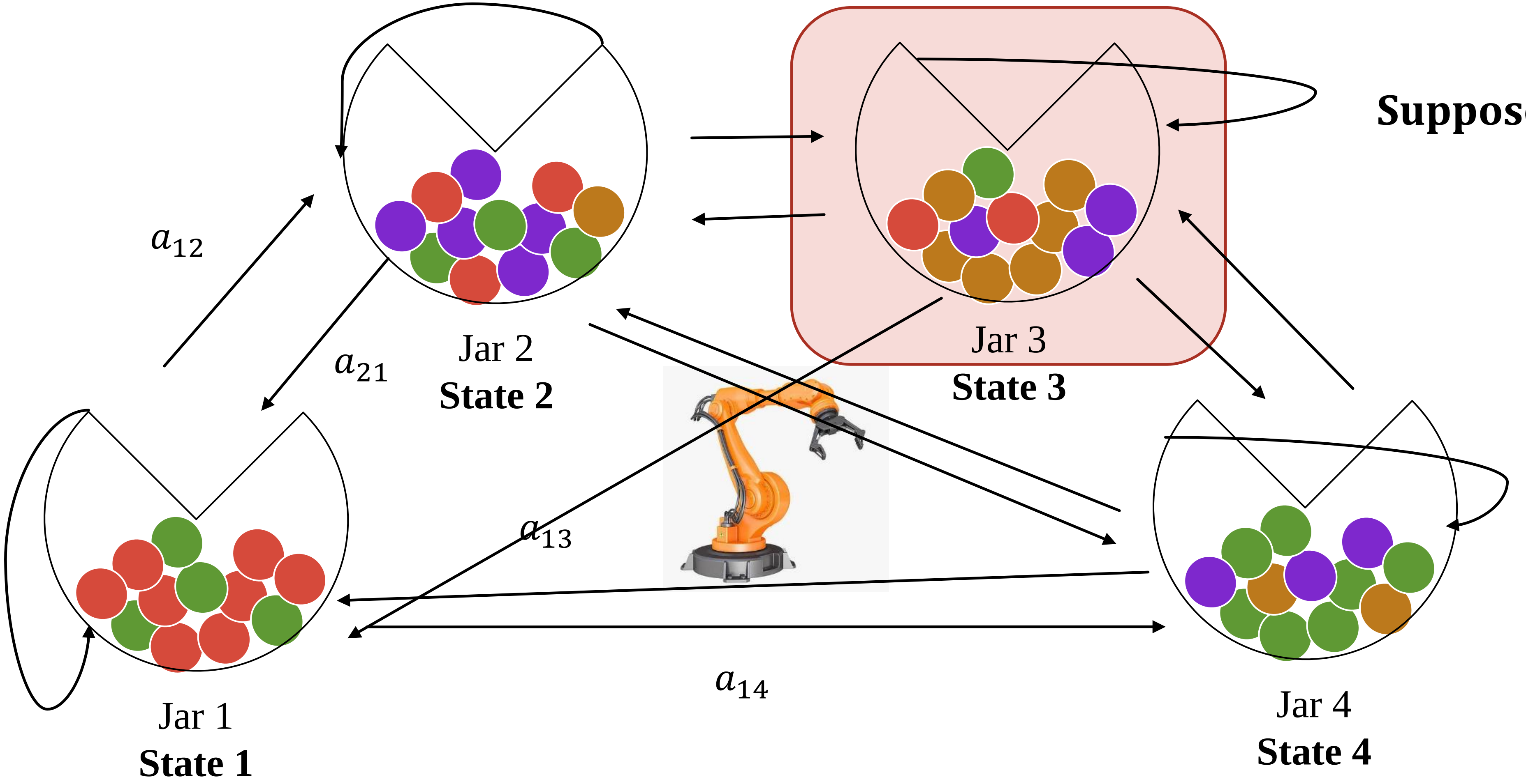
Problem with This Approach



Suppose $a_{13} = 0$

S2, S4, S1, S3

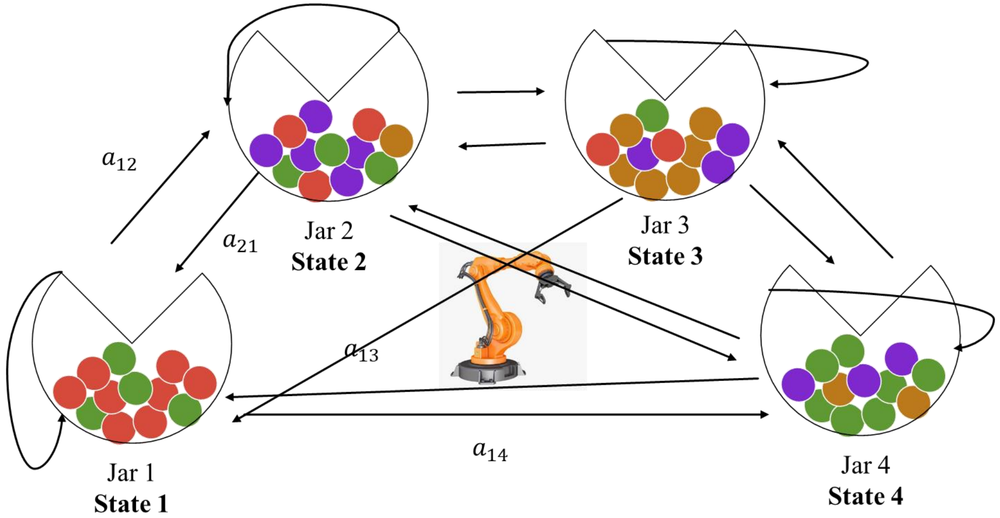
Problem with This Approach



Suppose $a_{13} = 0$

S2, S4, S1, ~~S3~~

Question 2



Choose a path sequence that maximizes

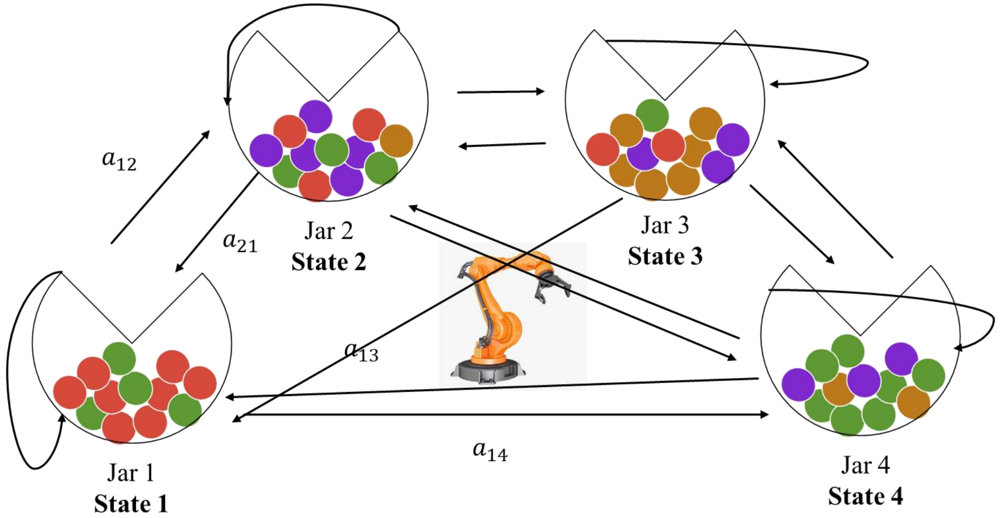
$$P(Q|O,\lambda)$$

- Suppose, we have an observation sequence

$$O = \{O_1, O_2, \ldots, O_T\}$$

- Given the model, what sequence of states **best** explains the above observation?
- What is the meaning of best?

Question 2



- Suppose, we have an observation sequence

$$O = \{O_1, O_2, \dots, O_T\}$$

- Given the model, what sequence of states **best** explains the above observation?

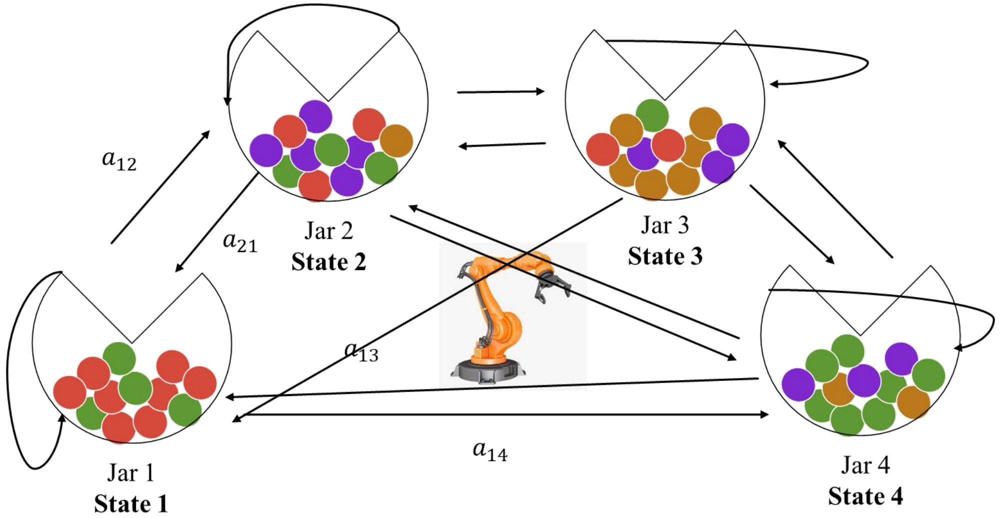
Choose a path sequence that maximizes

$$P(Q|O, \lambda)$$

This is equivalent to saying

$$P(Q, O|\lambda)$$

Question 2



- Suppose, we have an observation sequence

$$O = \{O_1, O_2, \ldots, O_T\}$$

- Given the model, what sequence of states **best** explains the above observation?

Choose a path sequence that maximizes

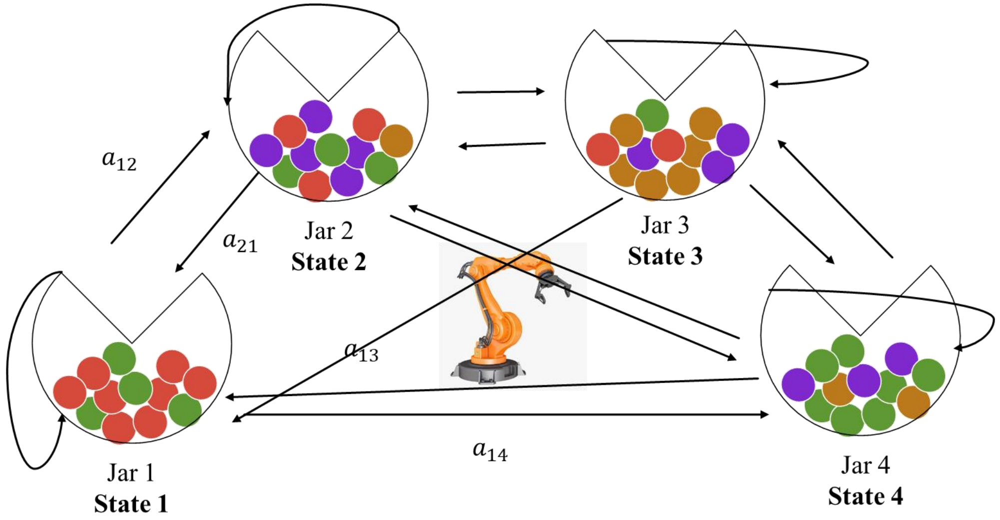
$$P(Q|O,\lambda)$$

This is equivalent to saying

$$P(Q,O|\lambda)$$

$$\boldsymbol{P(Q,O|\lambda) = P(\{q_1, q_2, \ldots, q_T\}, \{O_1, O_2, \ldots, O_T\}|\lambda)}$$

Solution: Viterbi Algorithm



- Suppose, we have an observation sequence

$$O = \{O_1, O_2, \dots, O_T\}$$

- Given the model, what sequence of states **best** explains the above observation?

Choose a path sequence that maximizes

$$P(Q|O, \lambda)$$

This is equivalent to saying

$$P(Q, O|\lambda)$$

$$\boldsymbol{P(Q, O|\lambda) = P(\{q_1, q_2, \dots, q_T\}, \{O_1, O_2, \dots, O_T\}|\lambda)}$$

Solution: Viterbi Algorithm

$$\boldsymbol{P(Q, O|\lambda) = P(\{q_1, q_2, \ldots, q_T\}, \{O_1, O_2, \ldots, O_T\}|\lambda)}$$

- To solve this, we need a variable

$$\delta_t(j) = \underbrace{max}_{q_1, q_2, \ldots, q_{t-1}} P(\{q_1, q_2, \ldots, q_t = j\}, \{O_1, O_2, \ldots, O_t\}|\lambda)$$

The highest probability of observing the sequence up to time t through a valid path that ends at state S_j

Solution: Viterbi Algorithm

- To solve this, we need a variable

$$\delta_t(j) = \underbrace{max}_{q_1,q_2,\ldots,q_{t-1}} P(\{q_1, q_2, \ldots, q_t = j\}, \{O_1, O_2, \ldots, O_t\}|\lambda)$$

The highest probability of observing the sequence up to time t through a valid path that ends at state S_j

- Induction step
  - $\delta_{t+1}(j) = [\underbrace{max}_{i} \{\delta_t(i)a_{ij}\}]b_j(O_{t+1})$

Solution: Viterbi Algorithm

- To solve this, we need a variable

$$\delta_t(j) = \underbrace{max}_{q_1,q_2,\dots,q_{t-1}} P(\{q_1, q_2, \dots, q_t = j\}, \{O_1, O_2, \dots, O_t\}|\lambda)$$

The highest probability of observing the sequence up to time t through a valid path that ends at state S_j

- Induction step
  - $\delta_{t+1}(j) = [\underbrace{max}_{i} \{\delta_t(i)a_{ij}\}]b_j(O_{t+1})$

- We have to keep track of where we came from (to do that, we will use another variable ψ)

Viterbi Algorithm

$$\delta_t(j) = \underbrace{max}_{q_1,q_2,\ldots,q_{t-1}} P(\{q_1, q_2, \ldots, q_t = j\}, \{O_1, O_2, \ldots, O_t\}|\lambda)$$

- Initialization
  - $\delta_1(i)$ =Highest probability of starting at state i and observing O_1

Viterbi Algorithm

$$\delta_t(j) = \underbrace{max}_{q_1,q_2,\dots,q_{t-1}} P(\{q_1, q_2, \dots, q_t = j\}, \{O_1, O_2, \dots, O_t\}|\lambda)$$

- Initialization
  - $\delta_1(i)$ =Highest probability of starting at state i and observing O_1
  - $\delta_1(i) = \pi_i b_i(O_1)$

Viterbi Algorithm

$$\delta_t(j) = \underbrace{max}_{q_1,q_2,\dots,q_{t-1}} P(\{q_1, q_2, \dots, q_t = j\}, \{O_1, O_2, \dots, O_t\}|\lambda)$$

- Initialization
  - $\delta_1(i)$ =Highest probability of starting at state i and observing O_1
  - $\delta_1(i) = \pi_i b_i(O_1)$
  - $\psi_1(i)$ indicates what is the path that we came through
  - $\psi_1(i) = 0$

Viterbi Algorithm

- Initialization
  - $\delta_1(i) = \pi_i b_i(O_1)$
  - $\psi_1(i) = 0$

- Inductive step
  - $\delta_t(j) = [\underbrace{max}_{1 \le i \le N} \{\delta_{t-1}(i) a_{ij}\}] b_j(O_t)$
  - $\psi_t(j) = \underbrace{argmax}_{1 \le i \le N} \{\delta_{t-1}(i) a_{ij}\}$

$2 \leq t \leq T$
$1 \leq j \leq N$

Viterbi Algorithm

- Initialization
  - $\delta_1(i) = \pi_i b_i(O_1)$
  - $\psi_1(i) = 0$
- Inductive step
  - $\delta_t(j) = [\underbrace{max}_{1 \leq i \leq N} \{\delta_{t-1}(i) a_{ij}\}] b_j(O_t)$
  - $\psi_t(j) = \underbrace{argmax}_{1 \leq i \leq N} \{\delta_{t-1}(i) a_{ij}\}$

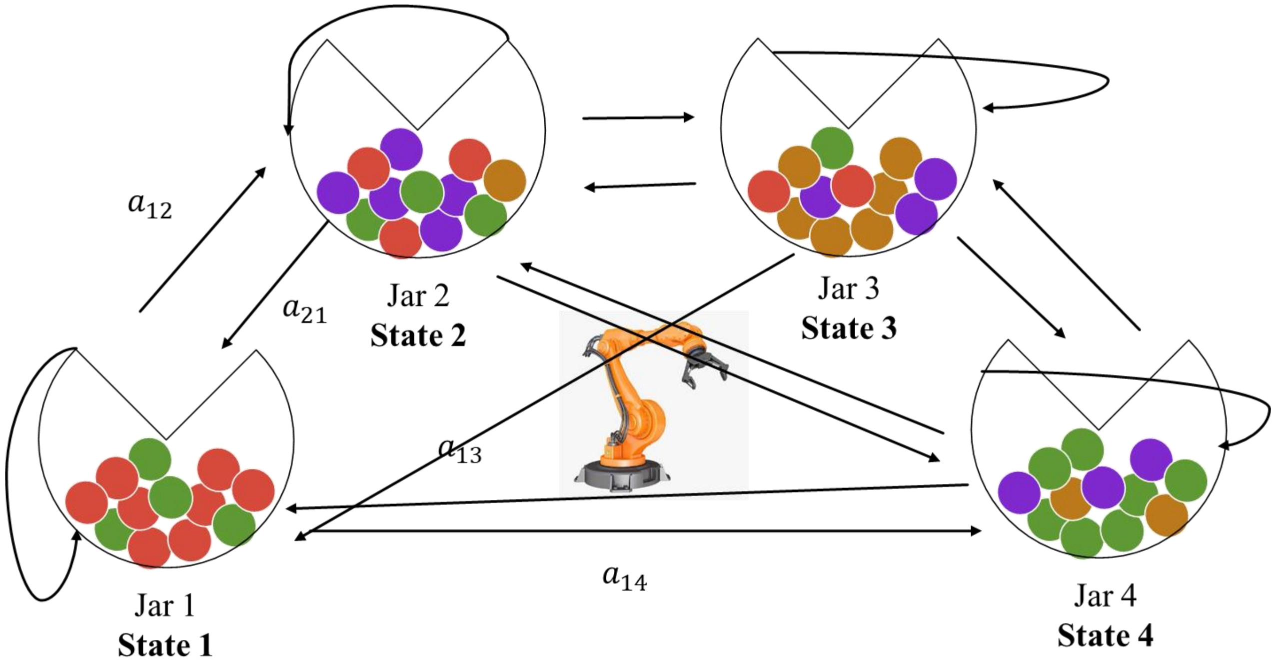
$2 \leq t \leq T$
$1 \leq j \leq N$

- Termination
  - $P^* = \underbrace{max}_{1 \leq i \leq N} \delta_T(i)$
  - $q_T^* = \underbrace{argmax}_{1 \leq i \leq N} \delta_T(i)$

Viterbi Algorithm

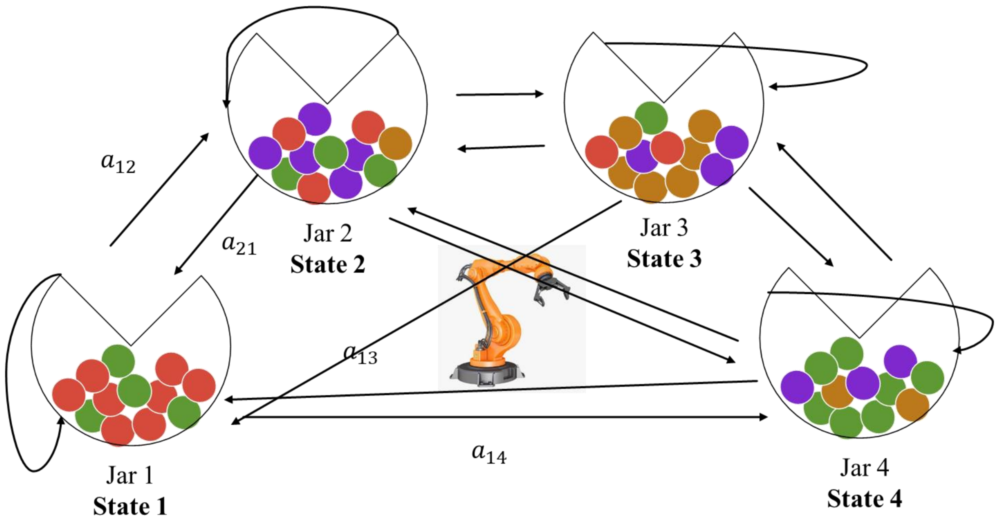
- Initialization
  - $\delta_1(i) = \pi_i b_i(O_1)$
  - $\psi_1(i) = 0$
- Inductive step
  - $\delta_t(j) = [\underbrace{max}_{1 \le i \le N} \{\delta_{t-1}(i) a_{ij}\}] b_j(O_t)$
  - $\psi_t(j) = \underbrace{argmax}_{1 \le i \le N} \{\delta_{t-1}(i) a_{ij}\}$

$$2 \le t \le T$$
$$1 \le j \le N$$



- Termination
  - $P^* = \underbrace{max}_{1 \le i \le N} \delta_T(i)$
  - $q_T^* = \underbrace{argmax}_{1 \le i \le N} \delta_T(i)$

Question 3

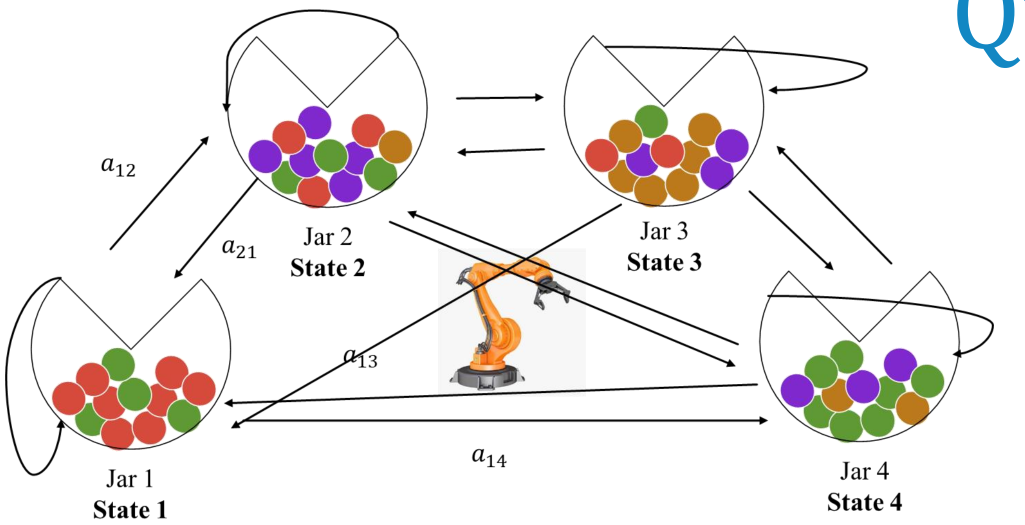


- Suppose, we have an observation sequence

  $$O = \{O_1, O_2, \ldots, O_T\}$$

- How can we find the model parameters $\lambda = \{A, B, \pi\}$ that would generate the above observation?

Question 3



- We have

$$\alpha_t(i) = P(O_1, O_2, \dots O_t, q_t = S_i | \lambda)$$

$$\beta_t(i) = P(O_{t+1}, O_{t+2}, \dots O_T | q_t = S_i, \lambda)$$

$$\gamma_t(i) = P(q_t = S_i | O, \lambda)$$

- Suppose, we have an observation sequence

$$O = \{O_1, O_2, \dots, O_T\}$$

- How can we find the model parameters $\lambda = \{A, B, \pi\}$ that would generate the above observation?

Baum-Welch Algorithm

- We have

$$\alpha_t(i) = P(O_1, O_2, \dots O_t, q_t = S_i | \lambda)$$

$$\beta_t(i) = P(O_{t+1}, O_{t+2}, \dots O_T | q_t = S_i, \lambda)$$

$$\gamma_t(i) = P(q_t = S_i | O, \lambda)$$

- Suppose, we have an observation sequence

$$O = \{O_1, O_2, \dots, O_T\}$$

- How can we find the model parameters $\lambda = \{A, B, \pi\}$ that would generate the above observation?

- **We can find locally optimal parameters (a good solution), but can't guarantee globally optimal parameters (the best result)**

Baum-Welch Algorithm

- We have

$$\alpha_t(i) = P(O_1, O_2, \ldots O_t, q_t = S_i|\lambda)$$

$$\beta_t(i) = P(O_{t+1}, O_{t+2}, \ldots O_T|q_t = S_i, \lambda)$$

$$\gamma_t(i) = P(q_t = S_i|O, \lambda)$$

- We need a new function

$$\boldsymbol{\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j|O, \lambda)}$$

Baum-Welch Algorithm

- We have

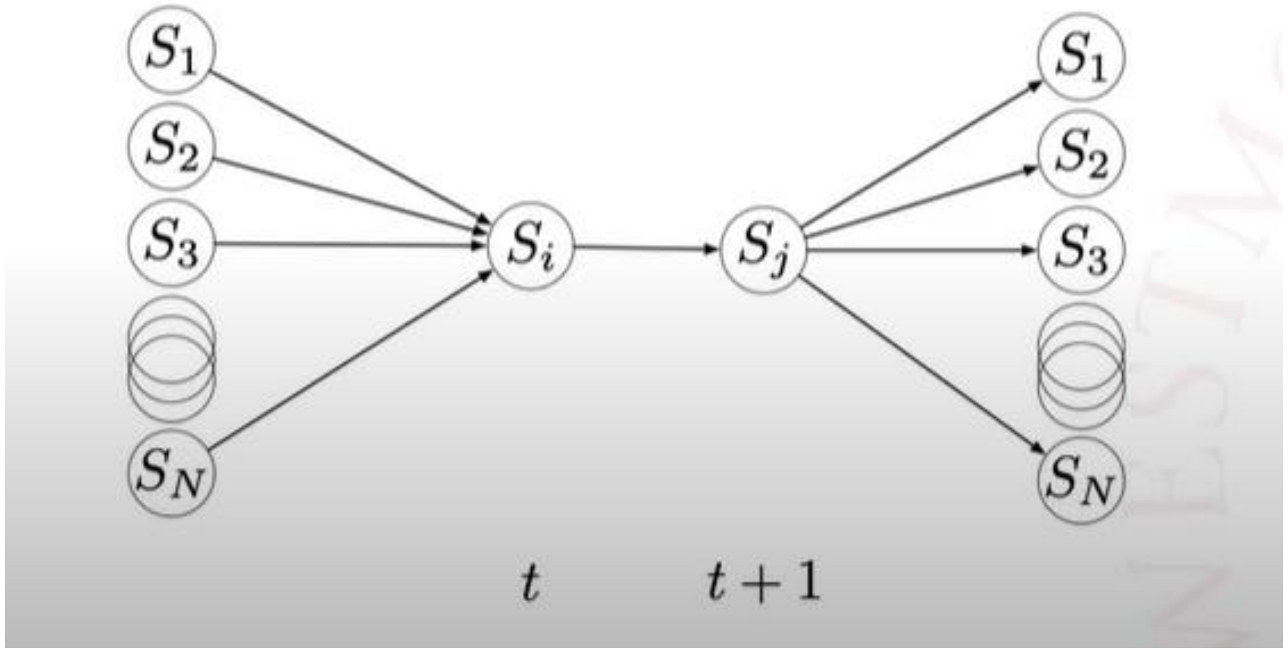
$$\alpha_t(i) = P(O_1, O_2, \dots O_t, q_t = S_i|\lambda)$$

$$\beta_t(i) = P(O_{t+1}, O_{t+2}, \dots O_T|q_t = S_i, \lambda)$$

$$\gamma_t(i) = P(q_t = S_i|O, \lambda)$$

- We need a new function

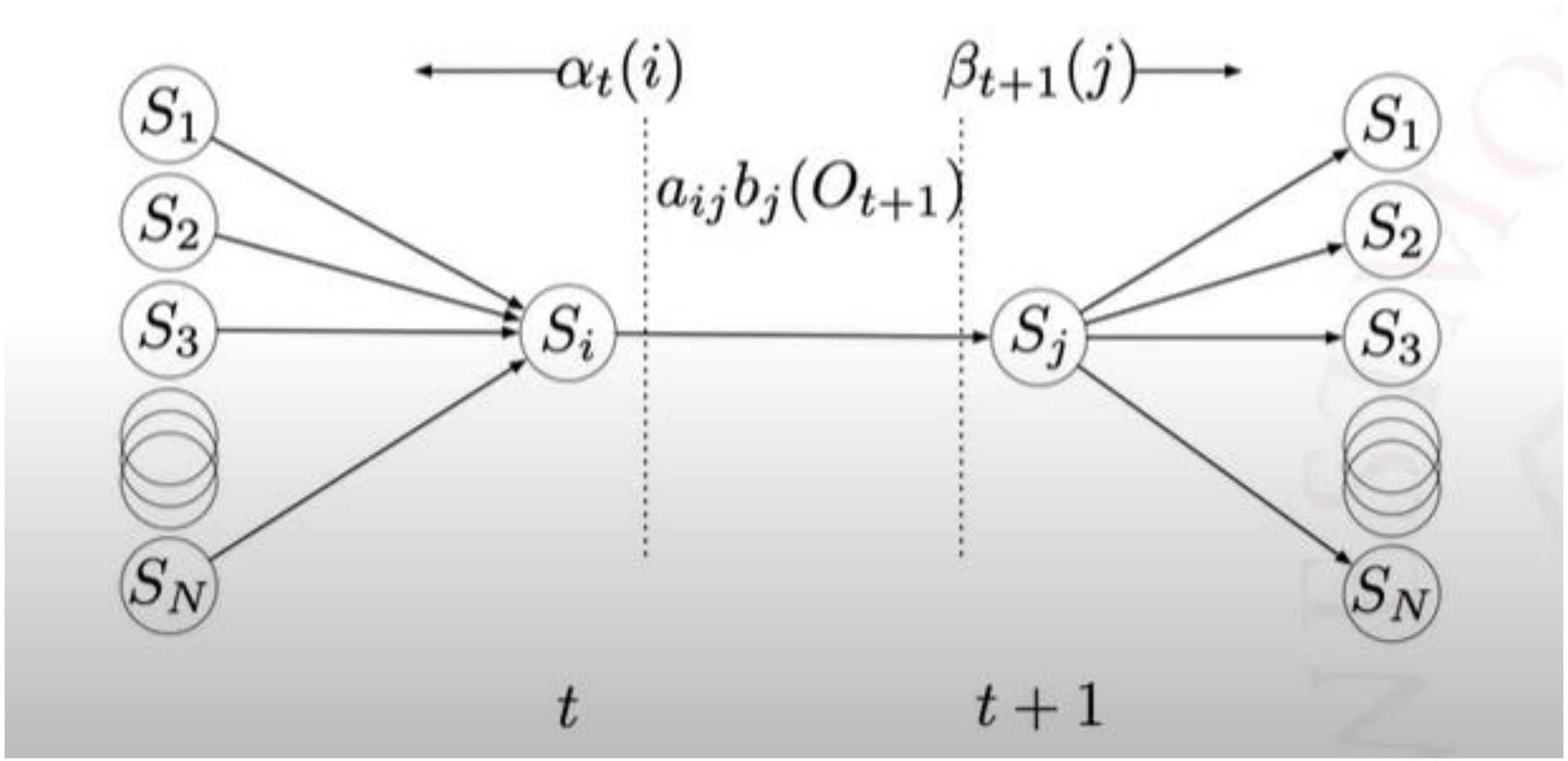
$$\boldsymbol{\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j|O, \lambda)}$$



Baum-Welch Algorithm

- We need a new function

$$\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j | O, \lambda)$$

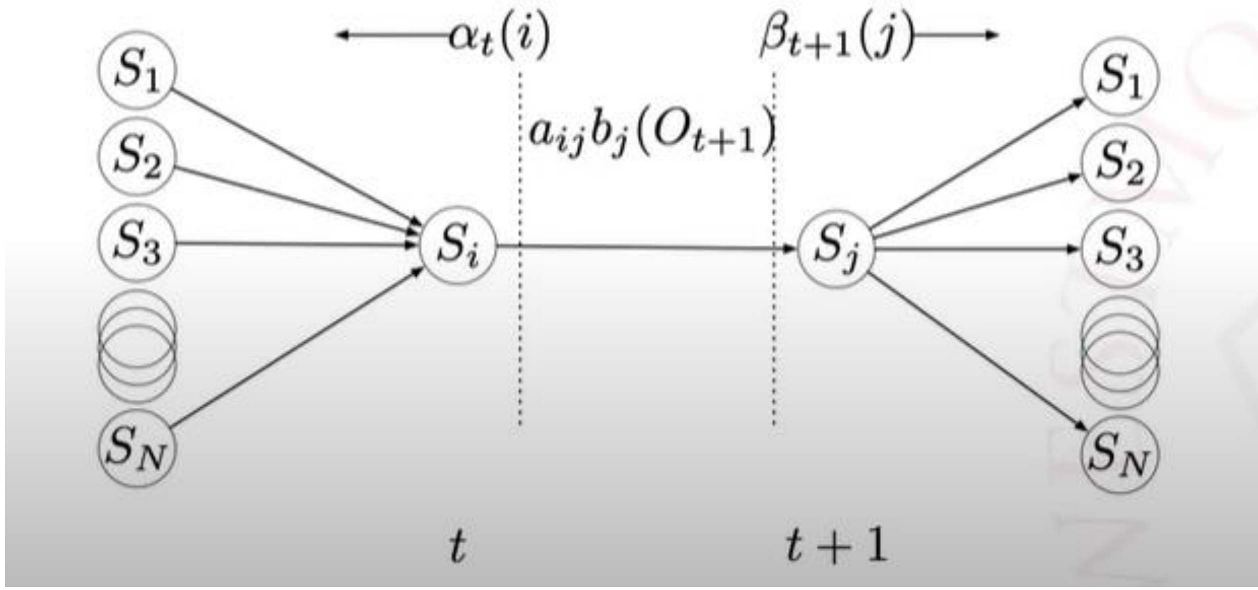


Baum-Welch Algorithm

- We need a new function

$$\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j | O, \lambda)$$

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{?}$$

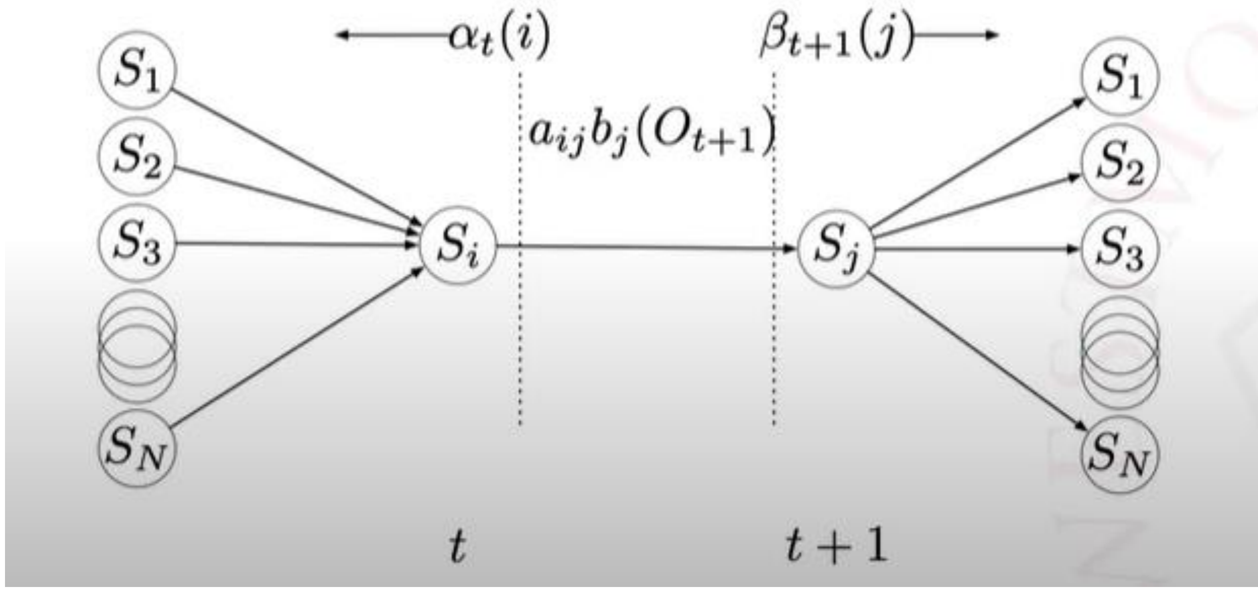


Baum-Welch Algorithm

- We need a new function

$$\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j | O, \lambda)$$

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O|\lambda)}$$

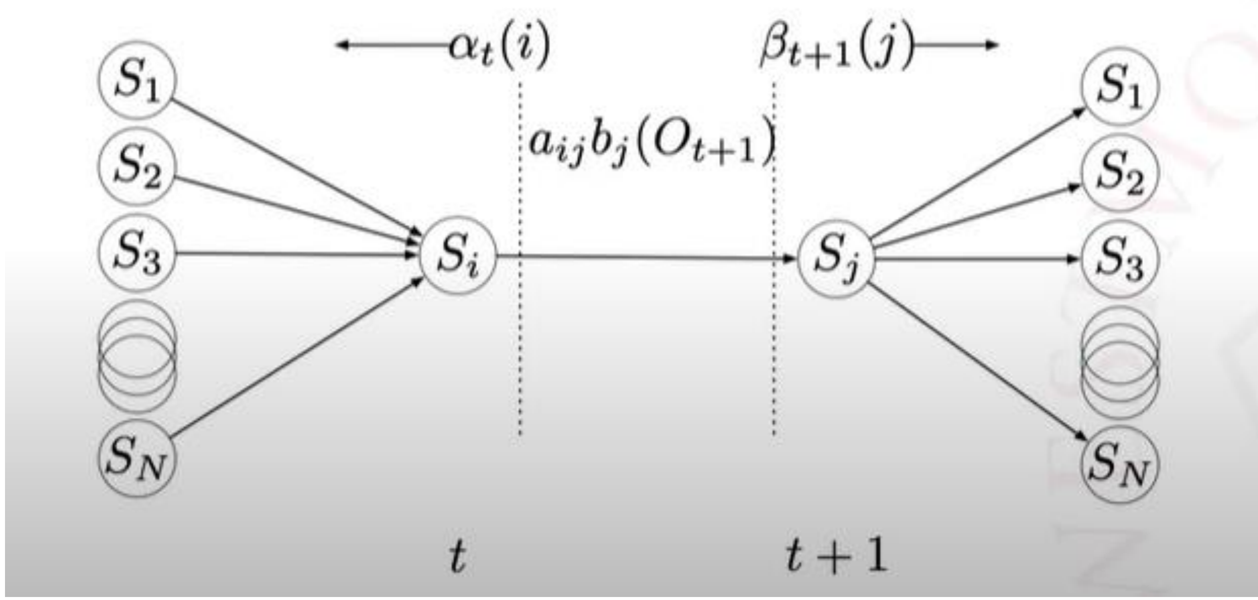


Baum-Welch Algorithm

- We need a new function

$$\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j | O, \lambda)$$

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O|\lambda)}$$



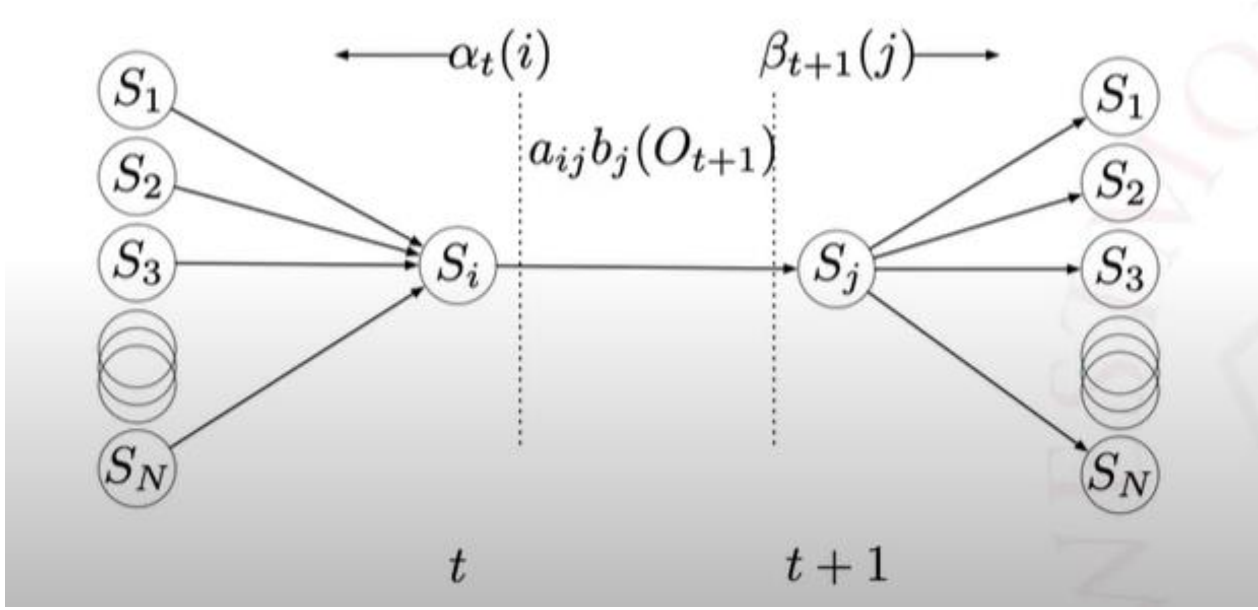
Baum-Welch Algorithm

- We need a new function

$$\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j | O, \lambda)$$

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O|\lambda)}$$

$$P(O|\lambda) = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)$$



Baum-Welch Algorithm

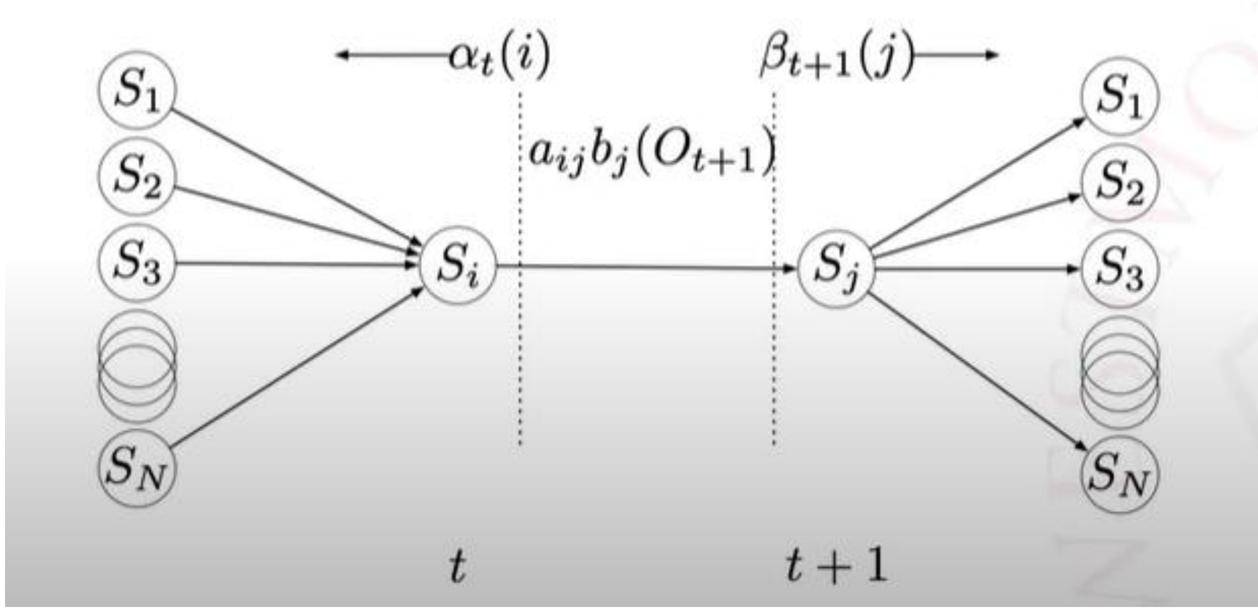
- We need a new function

$$\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j | O, \lambda)$$

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O|\lambda)}$$

$$P(O|\lambda) = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)$$

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}$$



Baum-Welch Algorithm

Randomly initialize A, B, π

Baum-Welch Algorithm

Randomly initialize A, B, π

Use Sequence of observations O

Baum-Welch Algorithm

Randomly initialize A, B, π

Use Sequence of observations O

Calculate $\alpha_t, \beta_t, \gamma_t, \xi_t$

Baum-Welch Algorithm

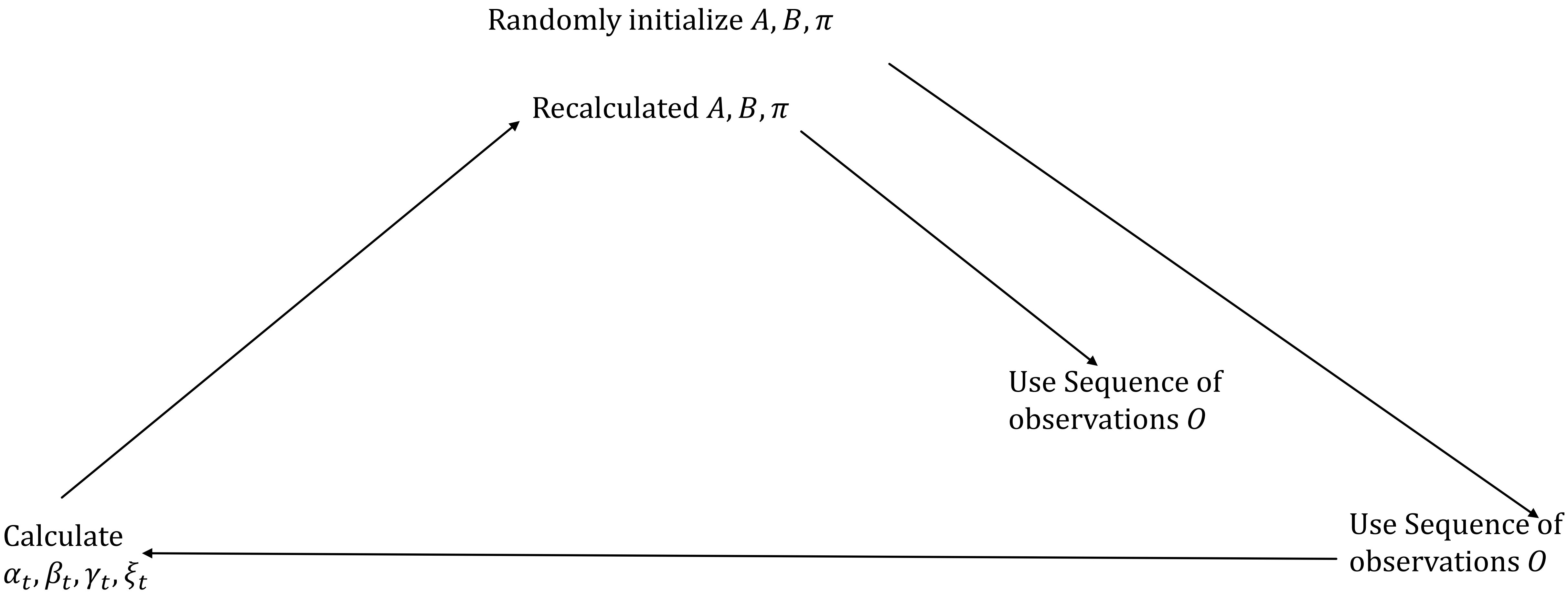
Randomly initialize A, B, π

Recalculated A, B, π

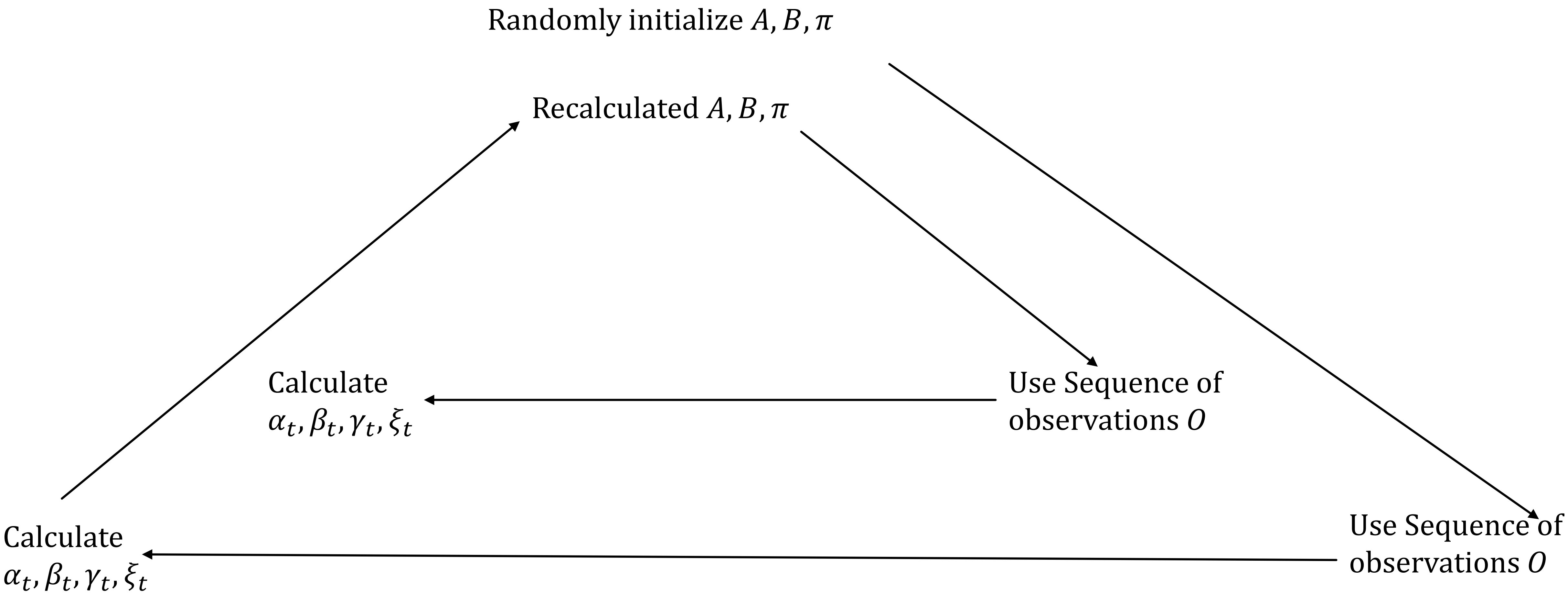
Calculate $\alpha_t, \beta_t, \gamma_t, \xi_t$

Use Sequence of observations O

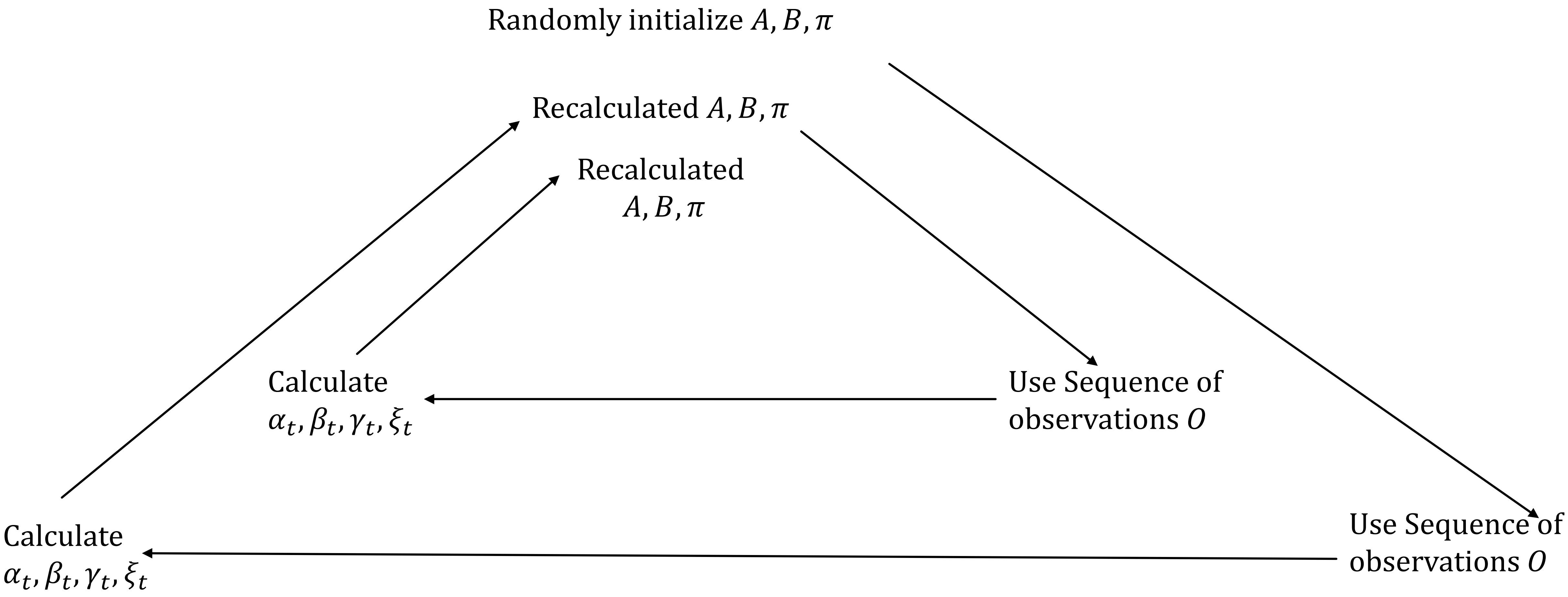
Baum-Welch Algorithm



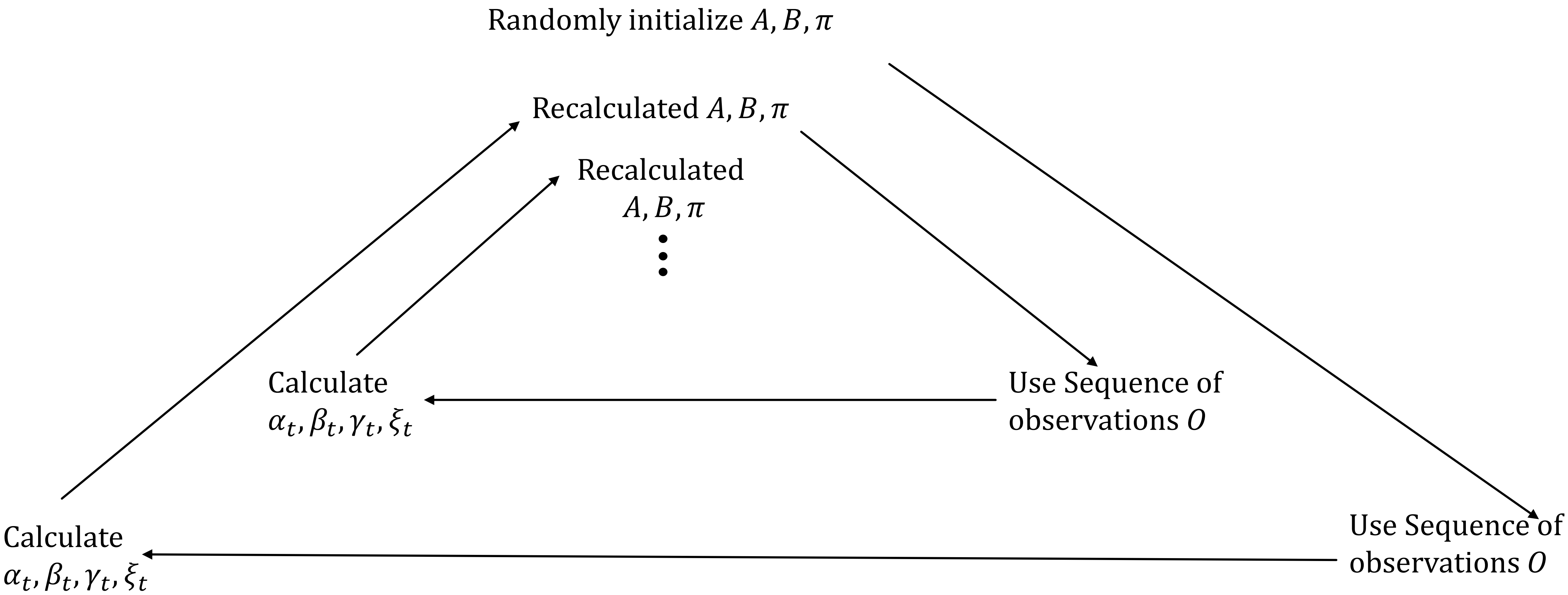
Baum-Welch Algorithm



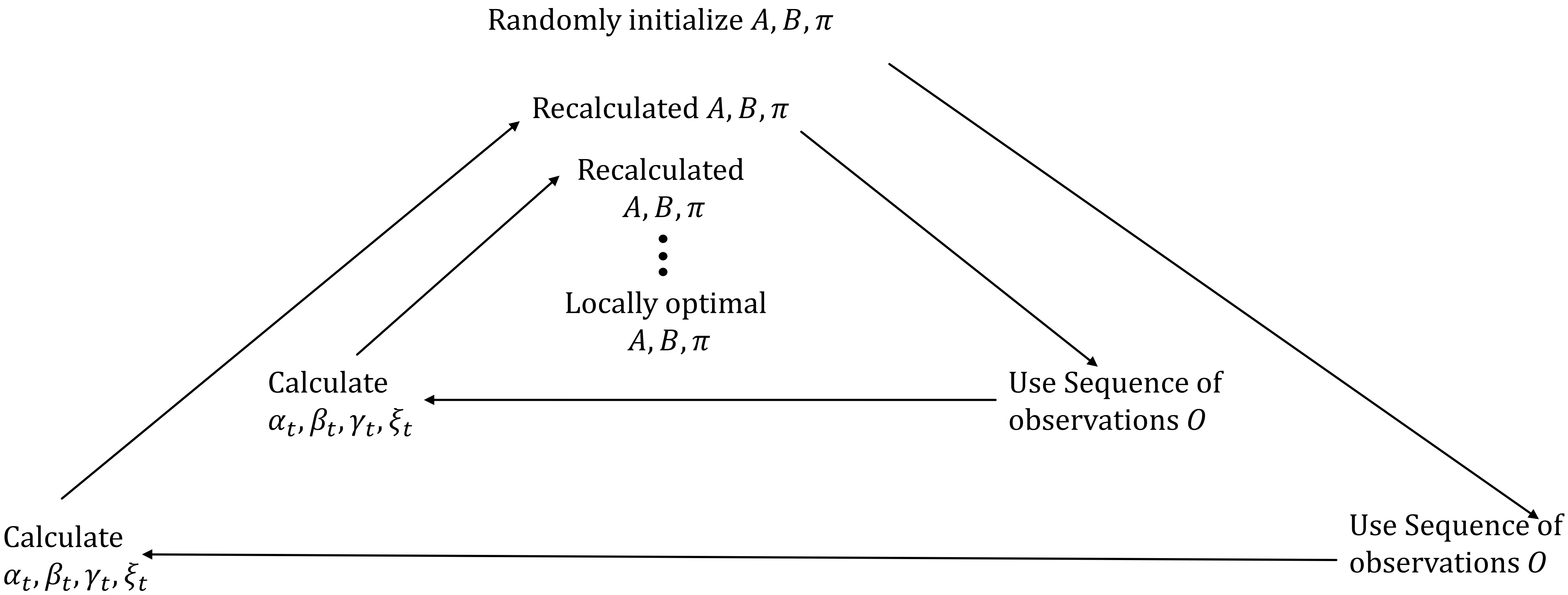
Baum-Welch Algorithm



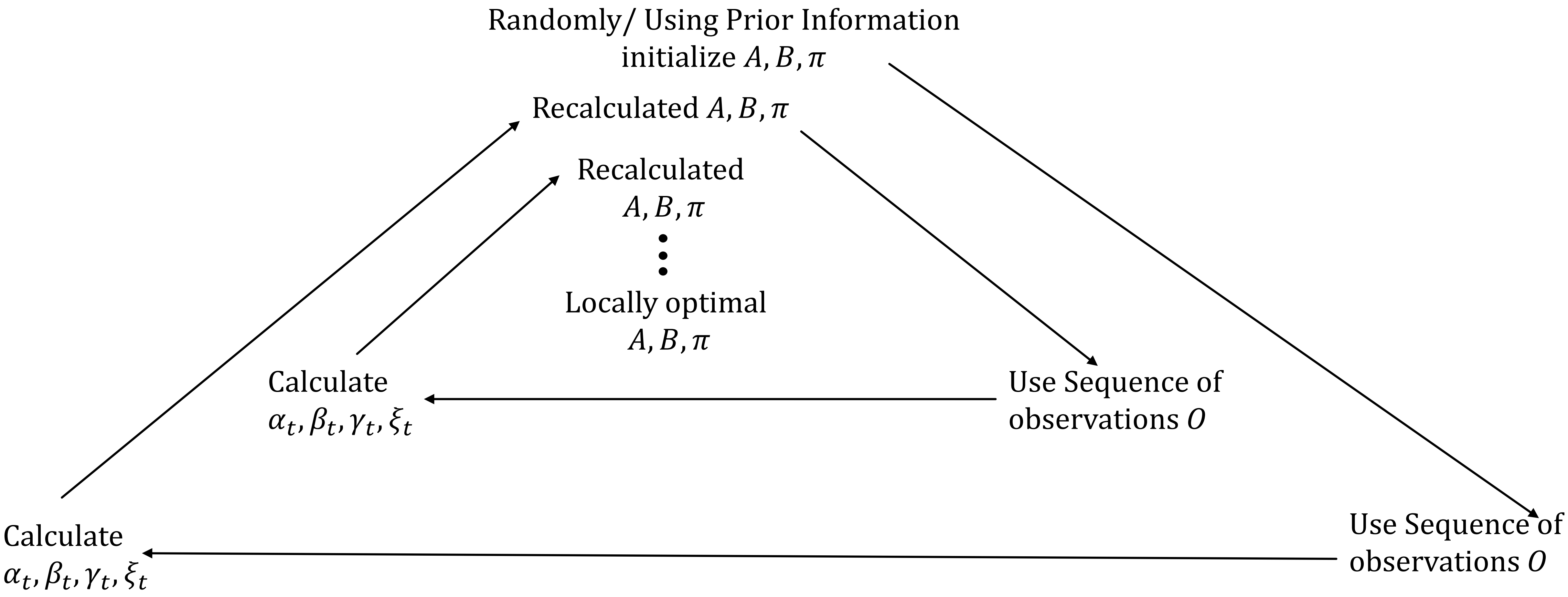
Baum-Welch Algorithm



Baum-Welch Algorithm



Baum-Welch Algorithm

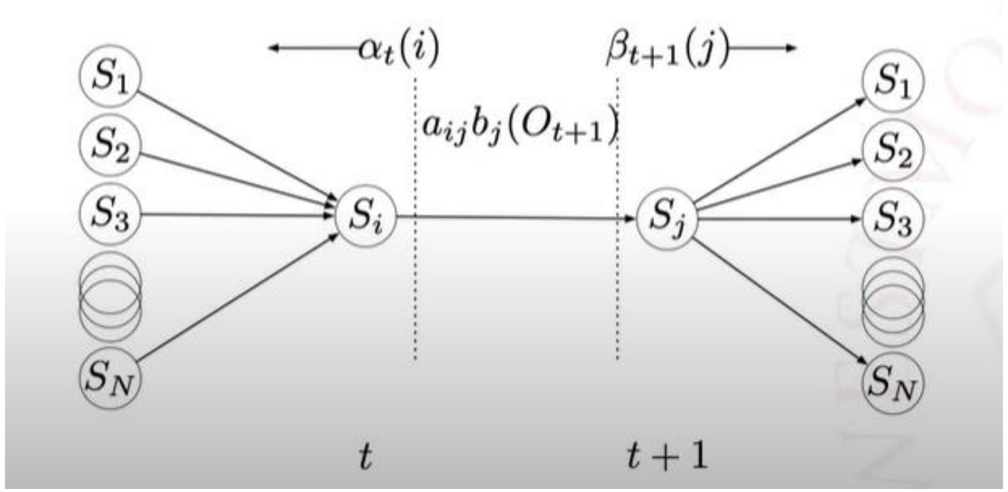


Finding Different Quantities

$$\boldsymbol{\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j | O, \lambda)}$$

- What is the probability of state i at time t given the observation

$$\gamma_t(i) = P(q_t = S_i | O, \lambda)$$



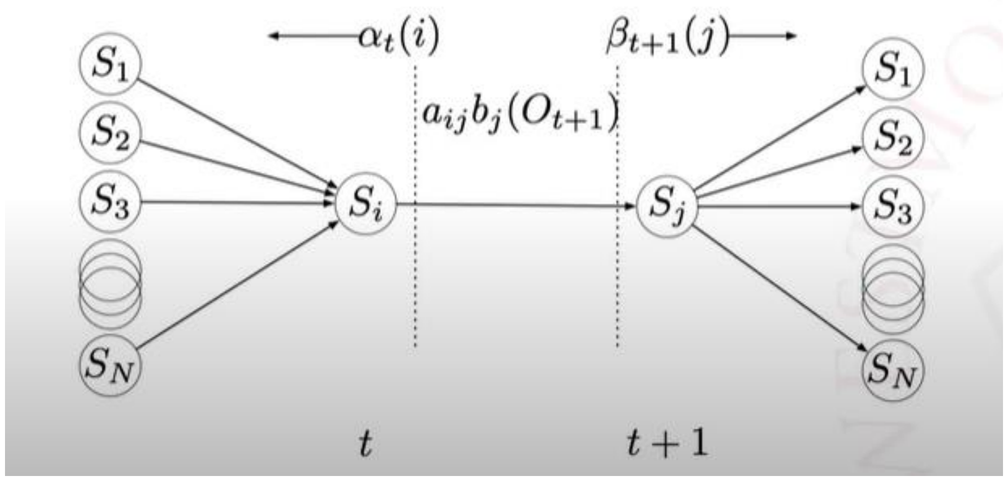
Finding Different Quantities

$$\boldsymbol{\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j | O, \lambda)}$$

- What is the probability of state i at time t given the observation

$$\gamma_t(i) = P(q_t = S_i | O, \lambda)$$

$$\boldsymbol{\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j)}$$



Finding Different Quantities

- Expected number of times S_i is ever visited and we moved out of S_i during our observations

$$\sum_{t=1}^{T-1} \gamma_t(i)$$

Finding Different Quantities

- Expected number of transitions from S_i to S_j during our observations

$$\sum_{t=1}^{T-1} \xi_t(i,j)$$

Finding Different Quantities

- Expected frequency of starting at state S_i

$$\boldsymbol{\gamma_1(i)}$$

Finding Different Quantities

- Expected frequency of starting at state S_i

$$\overline{\boldsymbol{\pi}_{\boldsymbol{i}}} = \boldsymbol{\gamma}_{\mathbf{1}}(\boldsymbol{i})$$

Finding Different Quantities

- Transition probability from S_i to S_j

$$\overline{a_{ij}} = \frac{expected\ number\ of\ transitions\ from\ S_i\ to\ S_j}{expected\ number\ of\ transitions\ from\ S_i}$$

$$= \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

Finding Different Quantities

- Observation probability at state S_i

$$\bar{b}_i(k) = \frac{expected\ number\ of\ times\ of\ being\ in\ state\ i\ and\ observing\ v_k}{expected\ number\ of\ times\ of\ being\ in\ state\ i}$$

$$= \frac{\sum^{T}_{\substack{t=1 \\ s.t.\ O_t = v_k}} \gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)}$$

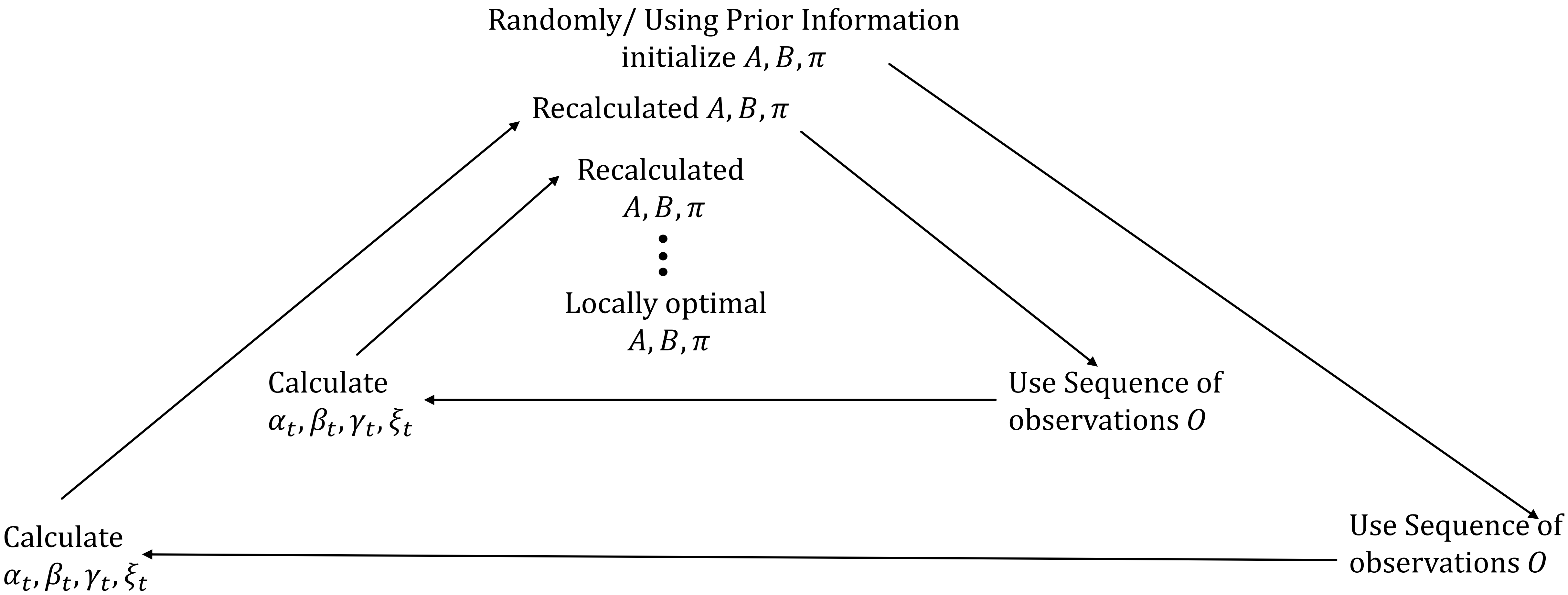
Finding Different Quantities

- So, we got new values $\overline{\boldsymbol{\pi_i}}$, $\overline{a_{ij}}$, $\overline{b}_i(k)$

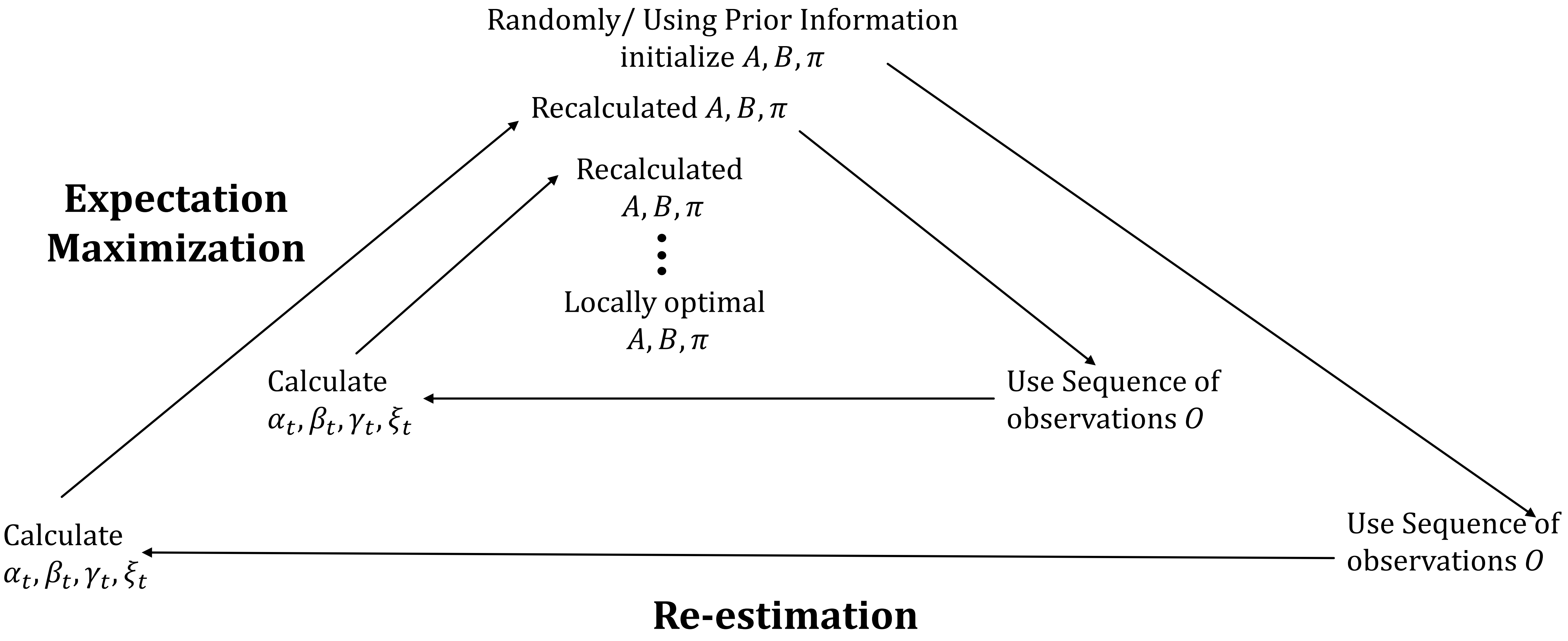
Finding Different Quantities

- So, we got new values $\overline{\boldsymbol{\pi_i}}$, $\overline{a_{ij}}$, $\overline{b}_i(k)$
- **That means we got a new model $\bar{\lambda}$**

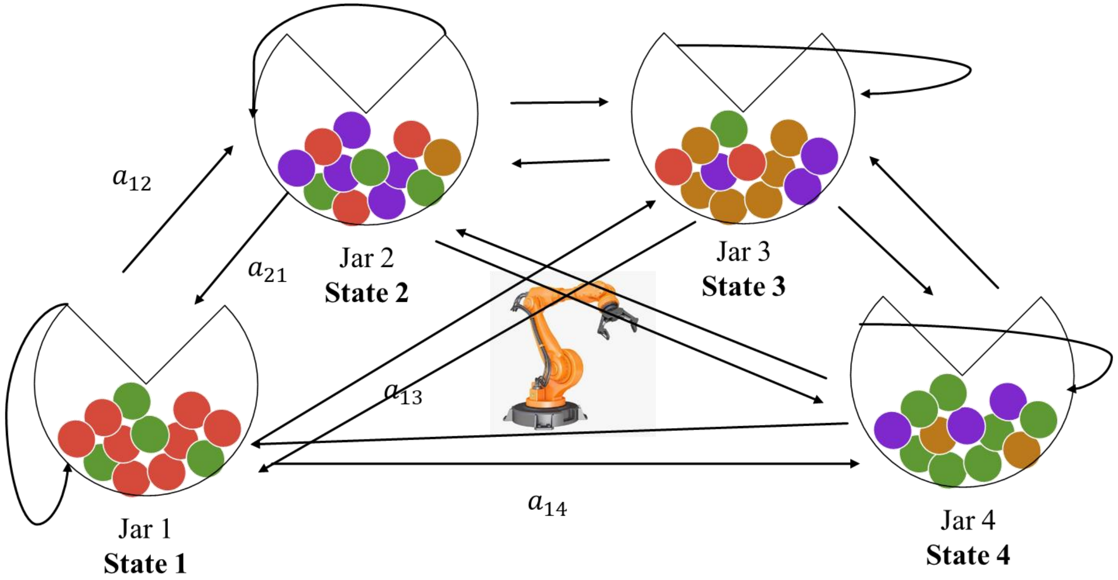
Baum-Welch Algorithm



Baum-Welch Algorithm



What Kind of Questions Can We Answer?



Given the model means given the information $\boldsymbol{\lambda = (A, B, \pi)}$

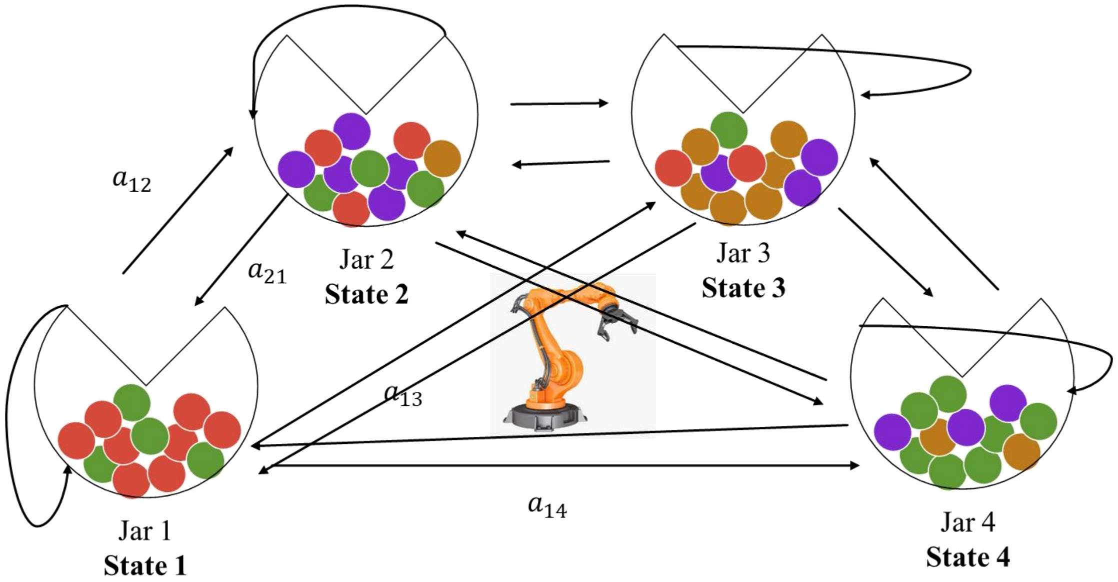
- Suppose, we have a model λ
- What is the probability that this model generates an observation sequence

$$O = \{O_1, O_2, \dots, O_T\}$$

for example, given our model, what is the probability of observing the following sequence

What is $\boldsymbol{P(O|\lambda)}$?

What Kind of Questions Can We Answer?



Given the information $\boldsymbol{\lambda = (A, B, \pi)}$

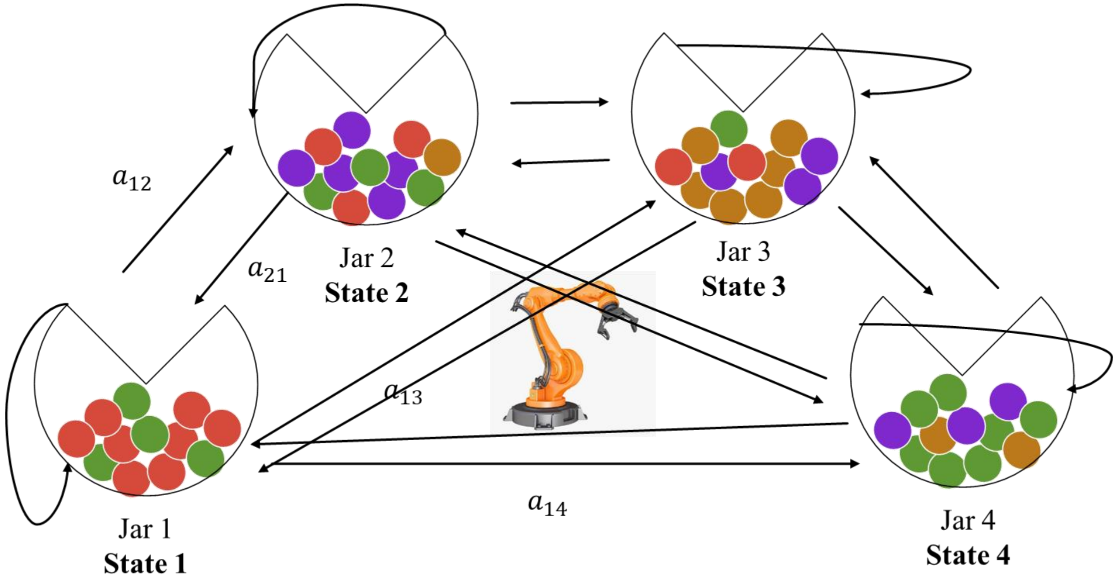
- Suppose, we have an observation sequence

$$O = \{O_1, O_2, \dots, O_T\}$$

- Given the model, what sequence of states best explains the above observation?

What is $\mathbf{Q = \{q_1, q_2, \dots, q_T\}}$?

What Kind of Questions Can We Answer?



Given $\boldsymbol{O} = \{\boldsymbol{O_1}, \boldsymbol{O_2}, \dots, \boldsymbol{O_T}\}$

What is $\boldsymbol{\lambda} = \{\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}\}$?

- Given an observation sequence

  $O = \{O_1, O_2, \dots, O_T\}$

- How to learn the model parameters that will maximize the chance of generating the above sequence?