Multi-Agent MDP with Controlled Observations

Jayadev Joy
Course Instructor: Dr. Quanyan Zhu

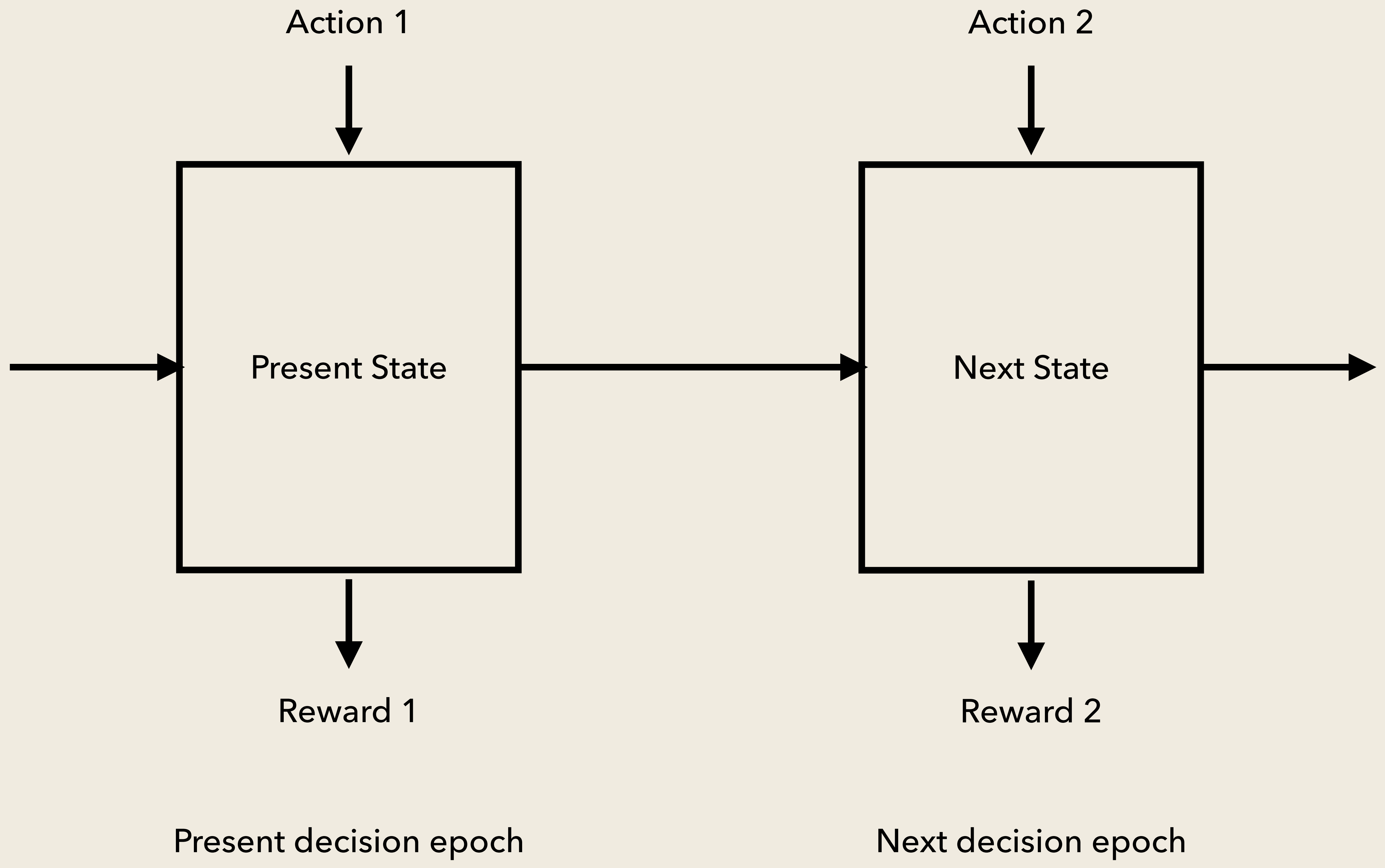
Department of Electrical and Computer Engineering, NYU Tandon

January 2023

NYU | TANDON SCHOOL OF ENGINEERING

Markov Decision Process

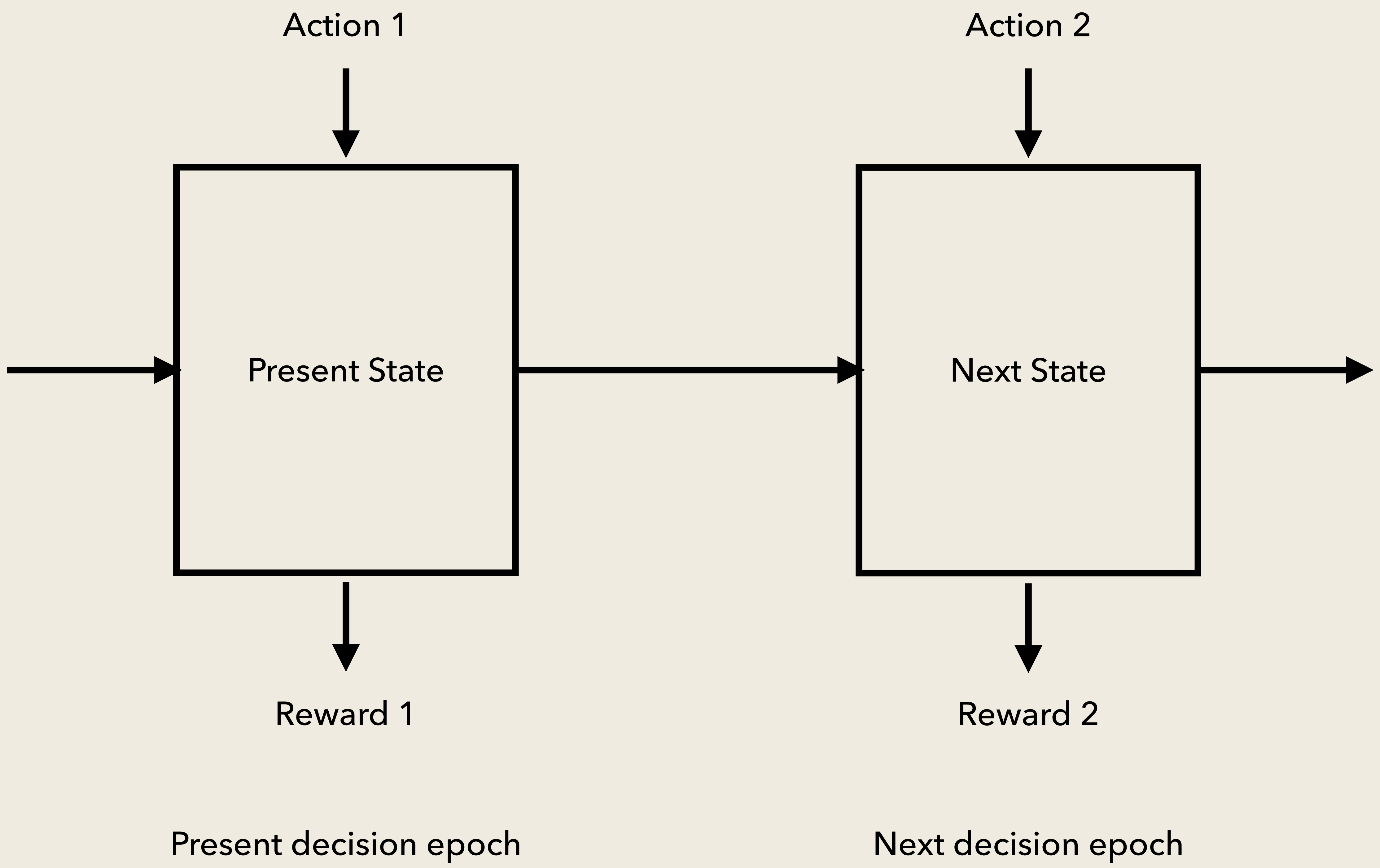
- A framework for sequential decision making in situations where the outcomes are partially random and partially under the control of a decision maker.
- Extension of Markov chains, with the addition of actions (introduces the element of control) and rewards (introduces the element of optimization).



Components of an MDP

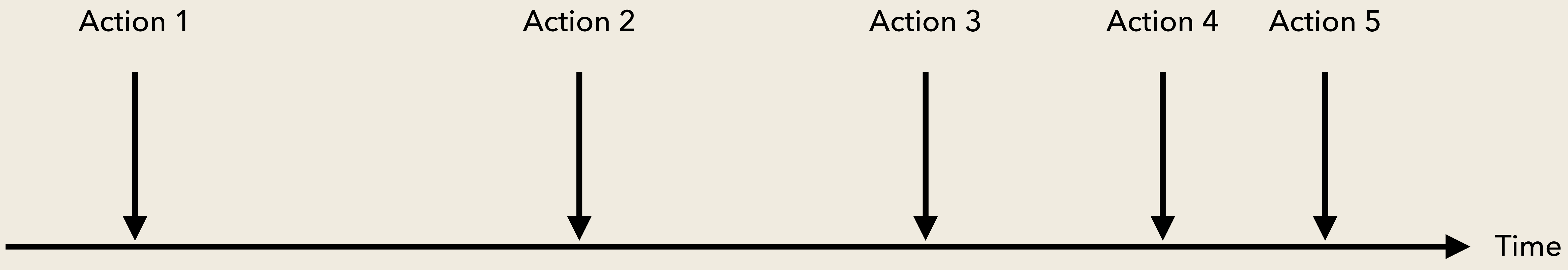
In this entire project we will be focusing on a continuous-time MDP.

- **State space** S
- **Action space** A
- **Transition Probability** $Q(x, x', a)$
- **Reward** $R(x, a)$
- **Policy function** π



Controlled Observations

- In some applications, continuous updates of observations could be limited or costly.
- Need for an MDP framework with controlled and limited observations.
- Decision maker determines the next observation time and the action for that interval.
- Joint determination of control trajectory and observation points.
- Introduces a trade-off between actions and observations.



[1] Yunhan Huang, Veeraruna Kavitha and Quanyan Zhu, "Continuous-Time Markov Decision Processes with Controlled Observations".

Notations (Single-Agent)

- Policy:

$$\pi(x) = \{(\alpha(x), \tau(x))\}$$

$$\alpha(x) = \{a(t) : a(\,.\,) \in L^p[t_x, t_x + T]\}; \tau(x) = \{T : T \in [\underline{T}, \infty]\}$$

- Transition probabilities:

$$q(x, x'; a, T) = P(X(T) = x' \,|\, X(0) = x, A([0,T]) = a(\,.\,), T_1 = T)$$

- Consolidated Rewards:

$$\begin{aligned}\bar{r}_k &= \bar{r}(X(\bar{T}_k), A(\bar{T}_k + \,.\,), T_k) \\ &= E\Big[\int_0^{T_k} \beta^t r(X(\bar{T}_k + t), A(\bar{T}_k + t) \,|\, X(\bar{T}_k), A(\bar{T}_k + \,.\,), T_k)\, dt\Big]\end{aligned}$$

- Here, $\bar{T}_k = \sum_{i<k} T_i$

Problem Formulation (Single-Agent)

- We wish to optimize the given accumulated reward:

$$v(x) = \sup_{\pi(x)} J(\pi(x), x)$$

$$J(\pi(x), x) = \sum_{k=1}^{\infty} E\left[\beta^{\bar{T}_k}\left(\bar{r}_k - g(T_k)\right)\right]$$

Discounted cost discrete-time MDP (Single-Agent)

- Discount factor:

$$\underline{\beta} = \beta^{\underline{T}}$$

- Markovian state:

$$Z_k = (X(\bar{T}_k), \tilde{T}_k), \text{ where } \tilde{T}_k = \bar{T}_k - (k-1)\underline{T}$$

- Markovian action:

$$A_k = (a_k(\,.\,), T_k)$$

- Running Reward:

$$R(Z_k, A_k) = \beta^{\tilde{T}_k}\Big(\bar{r}(X(\bar{T}_k), a(\,.\,), T_k) - g(T_k)\Big)$$

Discounted cost discrete-time MDP (Single-Agent)

- Th problem can now be restated as:

$$v(x) = \sup_{\pi(x)} J(\pi(x), x)$$

$$J(\pi(x), x) = \sum_{k=1}^{\infty} \underline{\beta}^{k-1} R(Z_k, A_k)$$

- The value function satisfies the following dynamic programming equation:

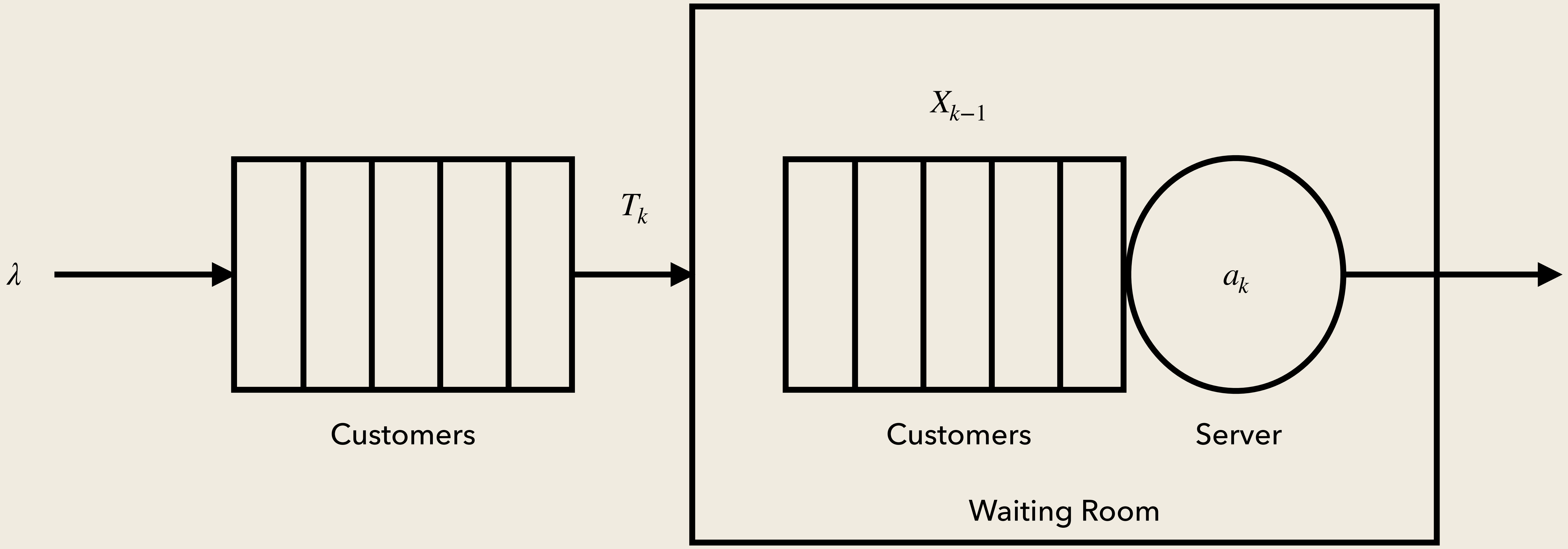
$$v(x) = \sup_{a,T} \left(\bar{r}(x, a(\,.\,), T) + \beta^{T} \sum_{x' \in S} q(x, x'; a, T) v(x') - g(T) \right)$$

- The optimal policy $\pi^*(x) = \{(\alpha^*(x), \tau^*(x))\} = \{(a^*(\,.\,), T^*)\}$, if it exists satisfies

$$v(x) = \bar{r}(x, a^*(\,.\,), T^*) + \beta^{T^*} \sum_{x' \in S} q(x, x'; a^*, T^*) v(x') - g(T^*) \text{ for all } x \in S$$

Gated Queuing Systems (Single Server)

- $\lambda =$ Arrival rate of customers (Poisson process)
- $W_k =$ Waiting time of all customers waiting during k^{th} observation period
- $T_k =$ Length of k^{th} observation period
- $X_{k-1} =$ Number of customers in inner room
- $a_k =$ Server speed



Gated Queuing Systems (Single Server)

- The cost here is the discounted expected waiting time:

$$\sum_{k=1}^{\infty} \beta^{\bar{T}_k} E[W_k]$$

- For a given X_{k-1}, T_k, and a_k, we have:

$$E[W_k \,|\, X_{k-1}, T_k, A_k = a_k] = \frac{\lambda T_k^2}{2} + \frac{X_{k-1}^2 + X_{k-1}}{2a_k}$$

- Finally, considering the cost for observations and sever speed, we want to optimize:

$$J(\pi(x), x) = \sum_{k=1}^{\infty} E\left[\beta^{\bar{T}_k}\left(\frac{\lambda T_k^2}{2} + \frac{X_{k-1}^2 + X_{k-1}}{2a_k} + g(T_k) + \eta(A_k)\right)\right]$$

Gated Queuing Systems (Single Server)

- Here, the transition probabilities are independent of the action:

$$q(x, x'; a, T) = e^{-\lambda T}\frac{(\lambda T)^{x'}}{(x')\,!}$$

- Assume a linear server speed cost:

$$\eta(a) = \eta a$$

- The optimal policy for the above problem is:

$$A_x^* = \sqrt{\frac{x(x+1)}{2\eta}}$$

$$T_x^* = T^*$$

Inventory Control

- $\theta =$ Targeted Inventory level
- $\mu =$ Departure Poisson rate
- $a(\,.\,) =$ Arrival Poisson rate (Control)
- $X_k =$ Inventory at k^{th} observation
- $T_k =$ Length of k^{th} observation period
- $\mathscr{A}(t; a) =$ Number of arrivals by time t
- $\mathscr{D}(t) =$ Number of departures by time t
- $\mathscr{N}(t; a) =$ Total number of arrivals and departures by time t
- $\nu =$ Cost of accelerating the arrivals

Inventory Control

- The inventory level at any given time before the $(k+1)^{th}$ epoch can be expressed as:

$$X_k(t) = X_k + \mathscr{A}(t;a) - \mathscr{D}(t) = X_k + \sum_{i=1}^{\mathscr{N}(t;a)} \xi_i$$

- The overall utility depends on the cost spent on the acceleration process and the deviation from the targeted inventory θ which is:

$$\int_0^\infty \beta^t \{E[(X_k(t) - \theta)^2] + \nu a(t)\} dt$$

- The consolidated utility for this example turns out to be:

$$r(x, a(t), t) = (x - \theta + \bar{a}(t) - \mu t)^2 + \bar{a}(t) + \mu t + a(t)\nu$$

Inventory Control

- The transition probability given the control $a(\, . \,)$ and time period T would be:

$$q(x, x'; a, T) = \sum_{k=x'-x}^{\infty} \frac{e^{-\bar{a}T-\mu T}(\bar{a}T)^{k}(\mu T)^{k-x'+x}}{k!(k-x'+x)!} \qquad \text{if} \quad x' > x$$

$$q(x, x'; a, T) = \sum_{k=x-x'}^{\infty} \frac{e^{-\bar{a}T-\mu T}(\bar{a}T)^{k-x+x'}(\mu T)^{k}}{k!(k-x+x')!} \qquad \text{if} \quad x' \leq x$$

- In each iteration k, for a given x, we need to solve the following optimal control problem:

$$v_{k+1}(x) = \inf_{a \in L^{\infty}[0,T]} \int_{0}^{T} \beta^{t} r(x, a(t), t)dt + \beta^{T} \sum_{x'} q(x, x'; a, T)v(x') + g(T)$$

Inventory Control

- At iteration k, we need to solve the following optimal control problem for any given x:

$$\inf_{a \in L^\infty[0,T]} \int_0^T \beta^t \{(x - \theta + y(t) - \mu t)^2 + y(t) + \mu t + a(t)\nu\} dt + h(y(T), T)$$

- When the Poisson arrival process is homogeneous, the problem becomes a finite dimensional optimization problem with:

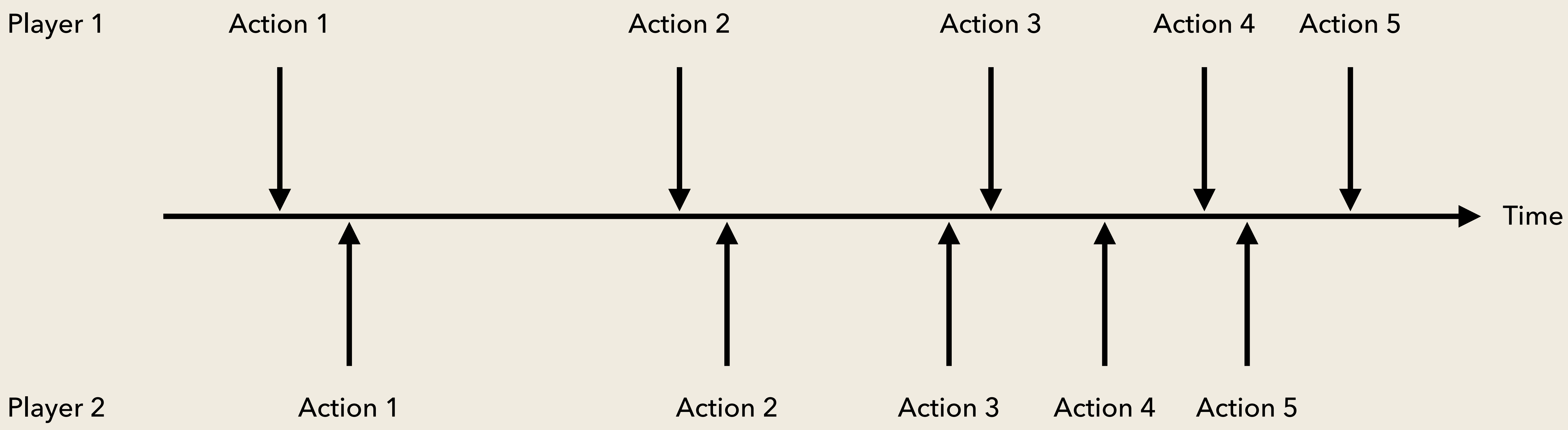
$$v_{k+1}(x) = \inf_{a \in [0,\bar{a}],T} \int_0^T \beta^t r(x, at, t) dt + h(aT, T)$$

- Hence, we have:

$$v_{k+1}(x) = \min_{a \in [0,\bar{a}],T} \int_0^T \beta^t \{(x - \theta + at - \mu)^2 + at + \mu t\} dt + h(aT, T)$$

Extension to Multi-Agent System

- In some scenarios, there will be more than one decision maker (player) in the system.
- Each player has to jointly determine the control trajectory and observation points.
- Introduces complexity since the system dynamics is governed by the actions of multiple players.
- Need for an MDP framework with controlled observations for multiple players.



Notations (Multi-Agent)

- Policy:

$$\pi^{(i)}(x) = \{(\alpha^{(i)}(x), \tau^{(i)}(x)) : \text{for any } x \in S\}$$

$$\alpha^{(i)}(x) = \{a^{(i)}(t) : a^{(i)}(\,.\,) \in L^p[t_x^{(i)}, t_x^{(i)} + T^{(i)}\}; \tau^{(i)}(x) = \{T^{(i)} : T^{(i)} \in [\underline{T}^{(i)}, \infty]\}$$

- Transition probabilities:

$$q(x, x'; a^{(1)}, a^{(2)}, T) = P(X(T) = x' \,|\, X(0) = x, A^{(1)}([0,T]) = a^{(1)}(\,.\,), A^{(2)}([0,T]) = a^{(2)}(\,.\,), T_1 = T)$$

Notations (Multi-Agent)

- Consolidated Rewards:

$$
\begin{aligned}
\bar{r}_k^{(i)} &= \bar{r}^{(i)}(X(\bar{T}_k^{(i)}), A^{(1)}(\bar{T}_k^{(i)} + .\,), A^{(2)}(\bar{T}_k^{(i)} + .\,), T_k^{(i)}) \\
&= E\Big[\int_0^{T_k^{(i)}} (\beta^{(i)})^t r(X(\bar{T}_k^{(i)} + t), A^{(1)}(\bar{T}_k^{(i)} + t), A^{(2)}(\bar{T}_k^{(i)} + t) \,|\, X(\bar{T}_k^{(i)}), a^{(i)}(\bar{T}_k^{(i)} + .\,), T_k^{(i)})\, dt \Big] rT_k), A(\bar{T}_k + .\,), T_k)\, dt \Big]
\end{aligned}
$$

- Here, $\bar{T}_k^{(i)} = \sum_{j<k} T_j^{(i)}$

Problem Formulation (Multi-Agent)

- We wish to optimize the given accumulated reward:

$$v^{(i)}(x) = \sup_{\pi^{(i)}(x)} J^{(i)}(\pi^{(1)}(x), \pi^{(2)}(x), x)$$

$$J^{(i)}(\pi^{(1)}(x), \pi^{(2)}(x), x) = \sum_{k=1}^{\infty} E\left[(\beta^{(i)})^{\bar{T}_k^{(i)}}\left(\bar{r}_k^{(i)} - g^{(i)}(T_k^{(i)})\right)\right]$$

Discounted cost discrete-time MDP (Multi-Agent)

- Discount factor:

$$\underline{\beta}^{(i)} = (\beta^{(i)})^{\underline{T}^{(i)}}$$

- Markovian state:

$$Z_k^{(i)} = (X(\bar{T}_k^{(i)}), \tilde{T}_k^{(i)}), \text{ where } \tilde{T}_k^{(i)} = \bar{T}_k^{(i)} - (k-1)\underline{T}^{(i)}$$

- Markovian action:

$$W_k^{(i)} = (a_k^{(1)}(\,.\,), a_k^{(2)}(\,.\,), T_k^{(i)})$$

- Running Reward:

$$R^{(i)}(Z_k^{(i)}, W_k^{(i)}) = (\beta^{(i)})^{\tilde{T}_k^{(i)}} \left(\bar{r}^{(i)}(X(\bar{T}_k^{(i)}), a^{(1)}(\,.\,), a^{(2)}(\,.\,), T_k^{(i)}) - g^{(i)}(T_k^{(i)}) \right)$$

Discounted cost discrete-time MDP (Multi-Agent)

- Th problem can now be restated as:

$$v^{(i)}(x) = \sup_{\pi^{(i)}(x)} J^{(i)}(\pi^{(1)}(x), \pi^{(2)}(x), x)$$

$$J^{(i)}(\pi^{(1)}(x), \pi^{(2)}(x), x) = \sum_{k=1}^{\infty} (\underline{\beta}^{(i)})^{k-1} R^{(i)}(Z_k^{(i)}, W_k^{(i)})$$

- The value function satisfies the following dynamic programming equation:

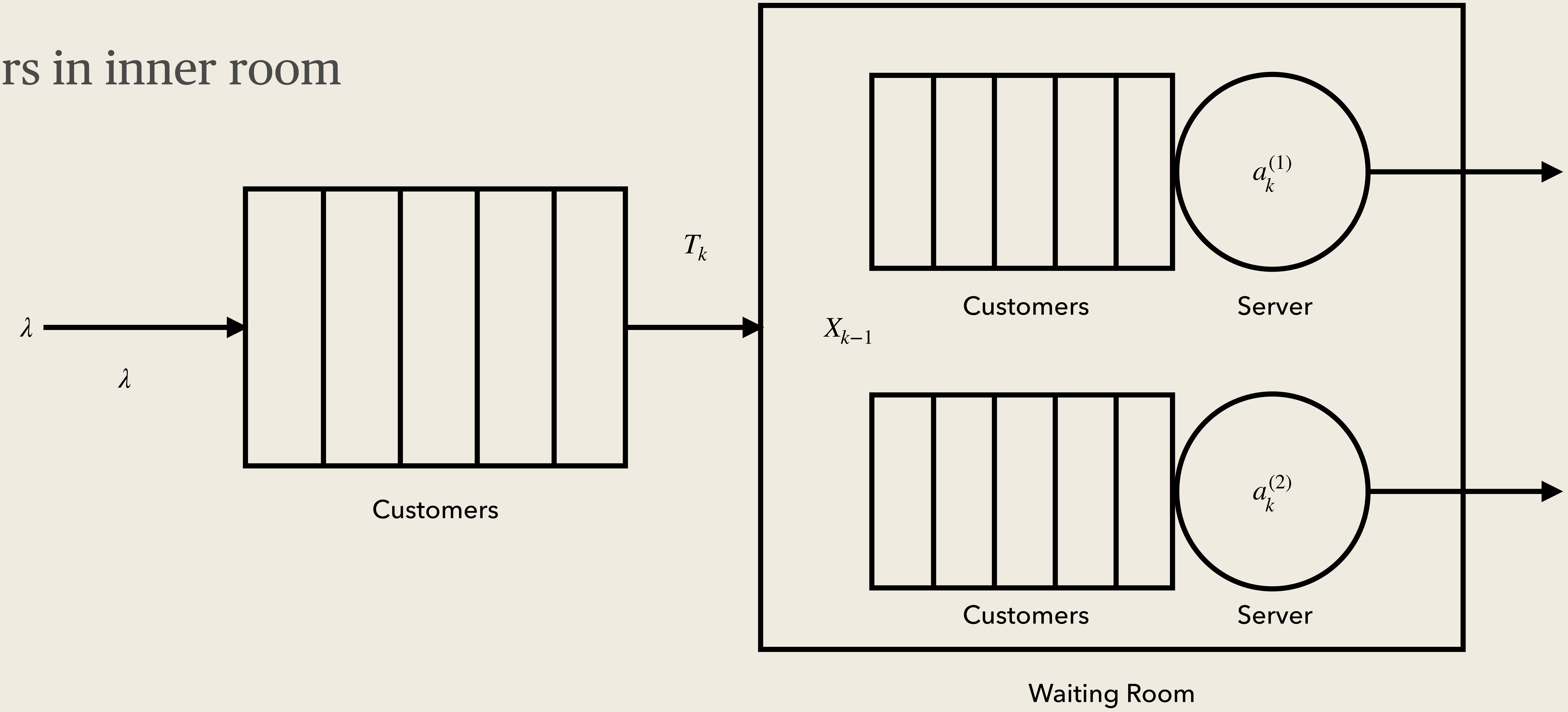
$$v^{(i)}(x) = \sup_{a^{(i)},T^{(i)}} \left(\bar{r}^{(i)}(x, a^{(1)}(\,.\,), a^{(2)}(\,.\,), T^{(i)}) + (\beta^{(i)})^{T^{(i)}} \sum_{x' \in S} q(x, x'; a^{(1)}, a^{(2)}, T^{(i)}) v^{(i)}(x') - g^{(i)}(T^{(i)}) \right)$$

- The optimal policy $\pi^{(i)*}(x) = \{(\alpha^{(i)*}(x), \tau^{(i)*}(x)) : x \in S\} = \{(a^{(i)*}(\,.\,), T^{(i)*})\}$, if it exists satisfies

$$v^{(i)}(x) = \bar{r}^{(i)}(x, a^{(i)*}(\,.\,), a^{-(i)}(\,.\,), T^{(i)*}) + (\beta^{(i)})^{T^{(i)*}} \sum_{x' \in S} q(x, x'; a^{(i)*}(\,.\,), a^{-(i)}(\,.\,), T^{(i)*}) v^{(i)}(x') - g^{(i)}(T^{(i)*})$$

Gated Queuing Systems (Multiple Servers)

- $\lambda =$ Arrival rate of customers (Poisson process)
- $W_k =$ Waiting time of all customers waiting during k^{th} observation period
- $T_k =$ Length of k^{th} observation period
- $X_{k-1} =$ Number of customers in inner room
- $a_k^{(1)} =$ Server speed 1
- $a_k^{(2)} =$ Server speed 2



Gated Queuing Systems (Multiple Servers)

- Objective 1 : We try to assign the server speeds for each server using X_{k-1}
- The cost here is the discounted expected waiting time:

$$\sum_{k=1}^{\infty} \beta^{\bar{T}_k} E[W_k]$$

- For a given X_{k-1}, T_k, $a_k^{(1)}$ and $a_k^{(2)}$, we have:

$$E[W_k \,|\, X_{k-1}, T_k, a_k^{(1)}, a_k^{(2)}] = \frac{\lambda T_k^2}{2} + \frac{X_{k-1}^2 + 2X_{k-1}}{2(a_k^{(1)} + a_k^{(2)})}$$

- Finally, considering the cost for observations and sever speed, we want to optimize:

$$J(\pi(x), x) = \sum_{k=1}^{\infty} E\left[\beta^{\bar{T}_k}\left(\frac{\lambda T_k^2}{2} + \frac{X_{k-1}^2 + 2X_{k-1}}{2(a_k^{(1)} + a_k^{(2)})} + g(T_k) + \eta^{(1)}(A_k^{(1)}) + \eta^{(2)}(A_k^{(2)})\right)\right]$$

Gated Queuing Systems (Multiple Servers)

- Here, the transition probabilities are independent of the action:

$$q(x, x'; a^{(1)}, a^{(2)}, T) = e^{-\lambda T} \frac{(\lambda T)^{x'}}{(x')!}$$

- Assume a linear server speed cost:

$$\eta^{(i)}(a^{(i)}) = \eta^{(i)} a^{(i)}$$

- The optimal policy for the above problem is to always assign the entire load to the server with the lower server speed cost and the corresponding speed and observation epochs are given by:

$$A_x^* = \sqrt{\frac{x(x+1)}{2\eta}}$$

$$T_x^* = T^*$$

Gated Queuing Systems (Multiple Servers)

- Objective 2 : We try to assign the number of customers for each server from which the server speed follows.
- The cost here is the discounted expected waiting time:

$$\sum_{k=1}^{\infty} \beta^{\bar{T}_k} E[W_k]$$

- For a given $X_{k-1}^{(1)}, X_{k-1}^{(2)}$, and T_k, and assuming a linear server speed cost we have:

$$E[W_k \,|\, X_{k-1}^{(1)}, X_{k-1}^{(1)}, T_k] = \frac{\lambda T_k^2}{2} + \sqrt{2\eta^{(1)} X_{k-1}^{(1)} (X_{k-1}^{(1)} + 1)} + \sqrt{2\eta^{(2)} X_{k-1}^{(2)} (X_{k-1}^{(2)} + 1)}$$

- Finally, considering the cost for observations, we want to optimize:

$$J(\pi(x), x) = \sum_{k=1}^{\infty} E\left[\beta^{\bar{T}_k} \left(\frac{\lambda T_k^2}{2} + \sqrt{2\eta^{(1)} X_{k-1}^{(1)} (X_{k-1}^{(1)} + 1)} + \sqrt{2\eta^{(2)} X_{k-1}^{(2)} (X_{k-1}^{(2)} + 1)} + g(T_k) \right)\right]$$

Gated Queuing Systems (Multiple Servers)

- Here, the transition probabilities are independent of the action:

$$q(x, x'; a^{(1)}, a^{(2)}, T) = e^{-\lambda T} \frac{(\lambda T)^{x'}}{(x')\,!}$$

- The optimal policy for the above problem is the solution to the given numerical expression and to assign the server speeds accordingly from $x^{(1)}$ and $x^{(2)}$:

$$\sqrt{\frac{\eta^{(1)}}{\eta^{(2)}}} = \frac{2x^{(2)} + 1}{2x^{(1)} + 1} \sqrt{\frac{x^{(1)}(x^{(1)} + 1)}{x^{(2)}(x^{(2)} + 1)}}$$

$$T_x^* = T^*$$

- The optimal server speeds for the servers follows directly from $x^{(1)}$ and $x^{(2)}$:

$$A_k^{(1)*} = \sqrt{\frac{x^{(1)}(x^{(1)} + 1)}{2\eta^{(1)}}}$$

Conclusion

- Provided a comprehensive summary of continuous-time MDP with controlled observations.
- Expanded this framework to encompass a multi-agent system setting.
- Considered two case studies for the single-agent MDP: Gated queuing system, and Inventory control.
- Extended our investigation to a multi-agent setting, specifically examining the gated queuing system using two distinct approaches.
- Assumed universal access to knowledge regarding the instantaneous server speeds among all servers.
- Future endeavors will delve into the analysis of a multiple-server gated queuing system, where centralized knowledge of instantaneous server speeds is not universally accessible.
- This analysis will necessitate the integration of a game-theoretic model into the MDP framework, adding a layer of complexity to the study.

Thank You