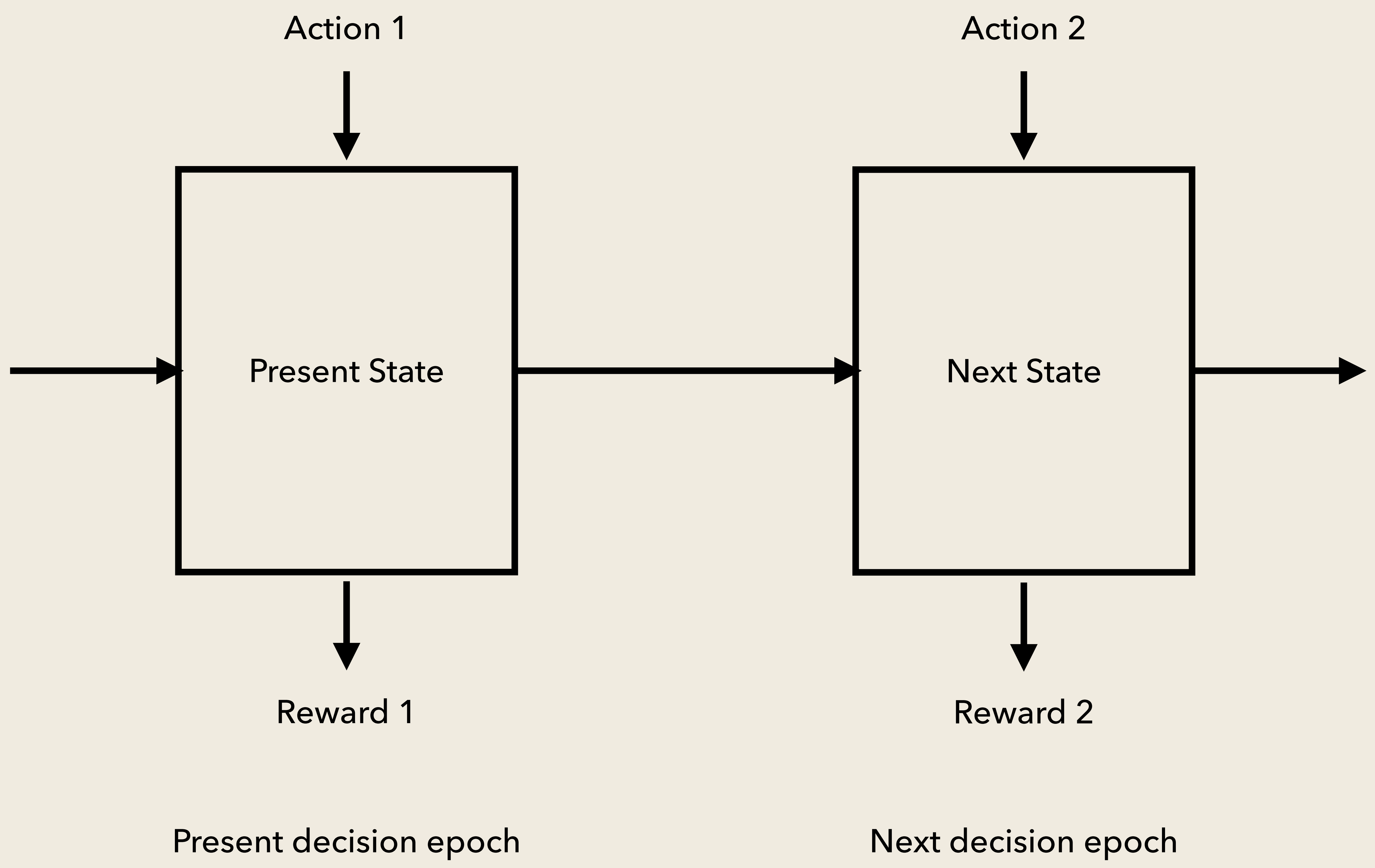
Multi-Agent MDP with Controlled Observations

Jayadev Joy
Course Instructor: Dr. Quanyan Zhu

Department of Electrical and Computer Engineering, NYU Tandon
October 2023

Markov Decision Process

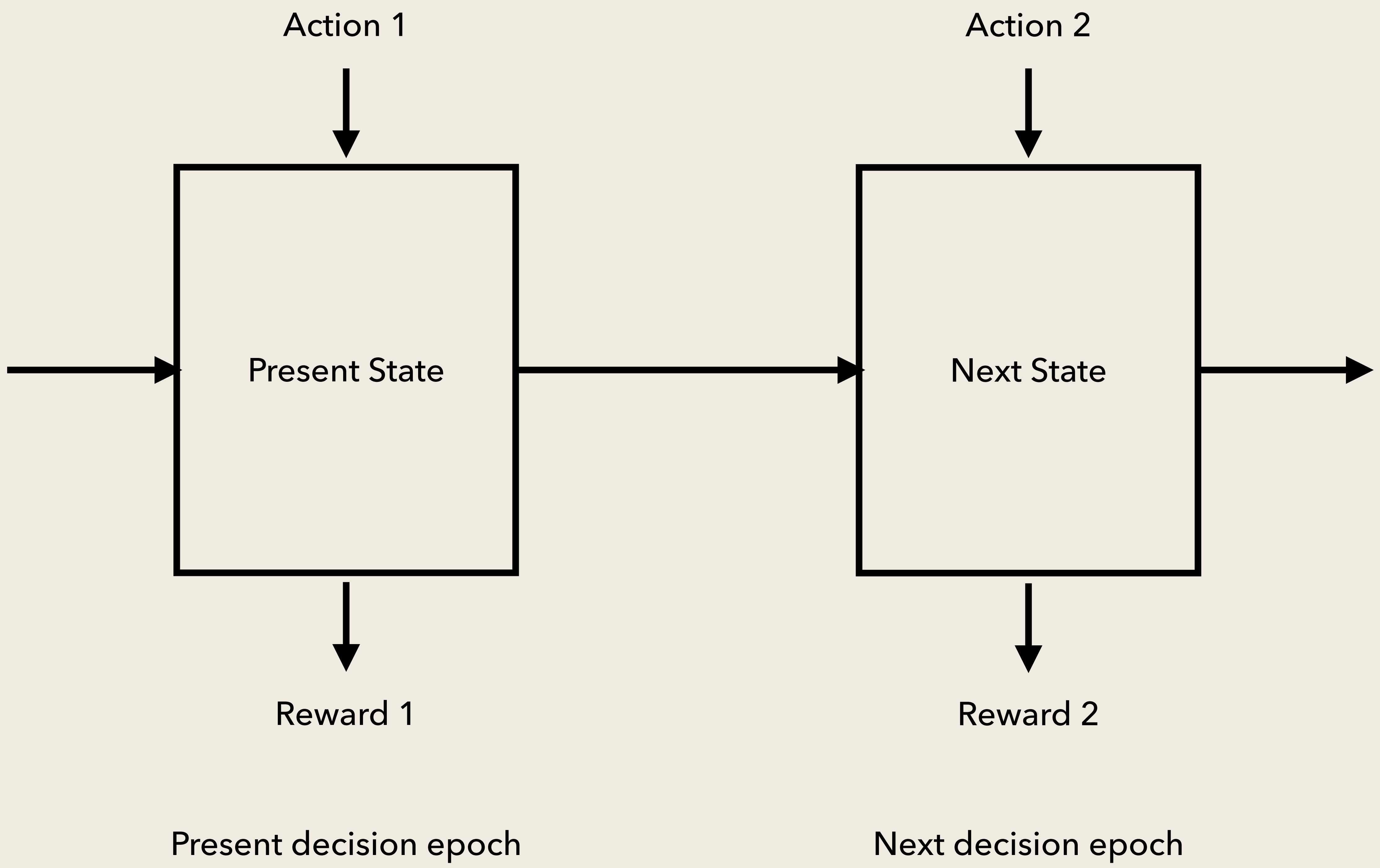
- A framework for sequential decision making in situations where the outcomes are partially random and partially under the control of a decision maker.
- Extension of Markov chains, with the addition of actions (introduces the element of control) and rewards (introduces the element of optimization).



Components of an MDP

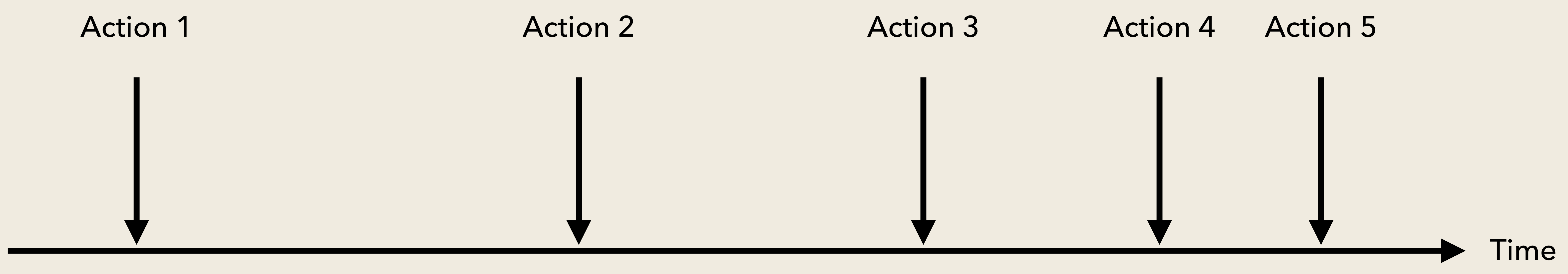
In this entire project we will be focusing on a continuous-time MDP.

- **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

- 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:

$$\bar{r}_k = \bar{r}(X(\bar{T}_k), A(\bar{T}_k + \,.\,), T_k)$$

$$= E\left[\int_0^{T_k} \beta^t r(X(\bar{T}_k + t), A(\bar{T}_k + t) \,|\, X(\bar{T}_k), A(\bar{T}_k + \,.\,), T_k)\, dt\right]$$

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

Problem Formulation

- 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

- 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

- 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:

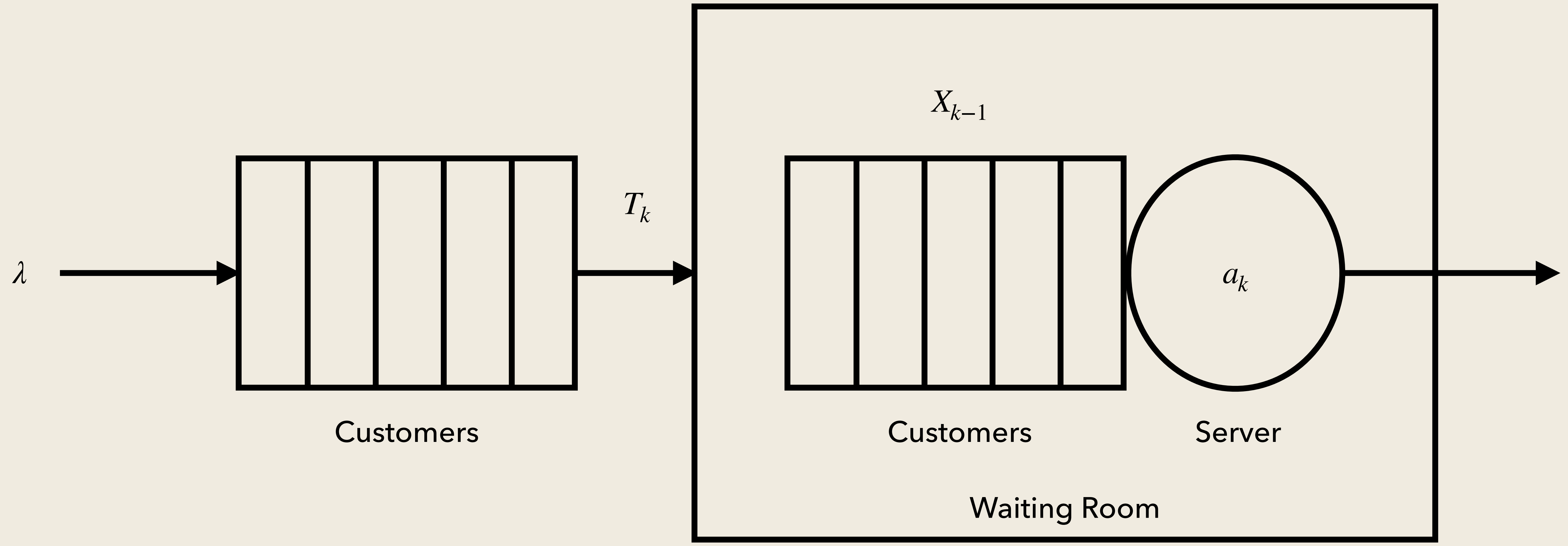
$$v(x) = \sup_{a,T} \left(\bar{r}(x, a(\,.\,), T) + \beta^T \sum_{x' \in S} q(x, x'; a, T) v(x') - g(T) \right)$$

- 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

- λ = 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

- 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

- 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:

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

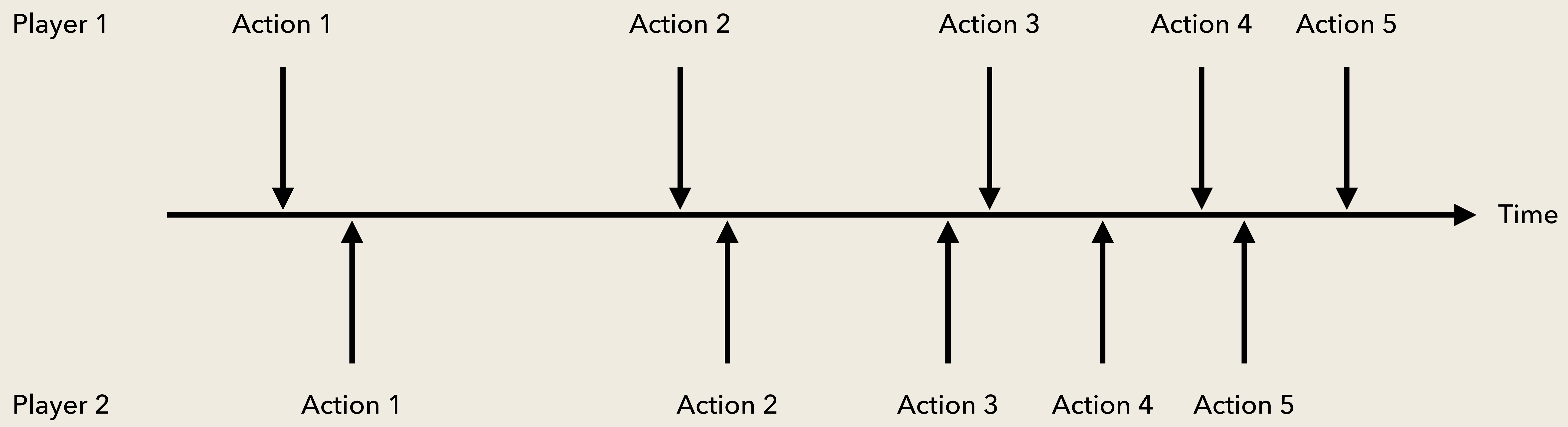
- The optimal policy for the above problem is:

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

$$T_x^* = T^*$$

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.



Extension to Multi-Agent System

- Policy:

$$\pi_i(x) = \{(\alpha_i(x), \tau_i(x))\} \text{ where } i \in \{a, b\}$$

$$\alpha_i(x) = \{i(t) : i(\,.\,) \in L^p[t_x, t_x + T_i]\};\ \tau_i(x) = \{T_i : T_i \in [\underline{T}_i, \infty]\}$$

- Transition probabilities:

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

- Consolidated Rewards:

$$\bar{r}_{ik} = \bar{r}(X(\bar{T}_{ik}), A(\bar{T}_{ik} + \,.\,), B(\bar{T}_{ik} + \,.\,), T_{ik})$$

$$= E\left[\int_0^{T_{ik}} \beta^t r(X(\bar{T}_{ik} + t), A(\bar{T}_{ik} + t), B(\bar{T}_{ik} + t), \,|\, X(\bar{T}_{ik}), i(\bar{T}_{ik} + \,.\,), T_{ik})\, dt\right]$$

- Here, $\bar{T}_{ik} = \sum_{j<k} T_{ij}$

Extension to Multi-Agent System

- We wish to optimize the given accumulated rewards:

$$v_i(x) = \sup_{\pi_i(x)} J_i(\pi_a(x), \pi_b(x), x)$$

$$J_i(\pi_a(x), \pi_b(x), x) = \sum_{k=1}^{\infty} E\left[\beta^{\bar{T}_{ik}}\left(\bar{r}_{ik} - g(T_{ik})\right)\right]$$

Thank You