Seminar of Ushio Lab.

B4 Yuma Yamakura

Outline

- Introduction
  - Markov Decision Process (MDP)
  - Q-learning
  - Discrete Event Systems (DESs)
  - Supervisory control
- Decentralized supervisory control of DESs based on reinforcement learning
- Simulation
- Future Work

Markov decision process (MDP)

Reference
木村元 ,“強化学習の基礎” ,計測と制御 第52巻 第1号 2013年1月号,公益社団法人 計測自動制御学会,2013

A finite MDP : $< X, U, P, R >$

X:the finite set of environment states

A:the finite set of agent actions

$P: X \times A \times X \rightarrow [0,1]$ the transition probability function

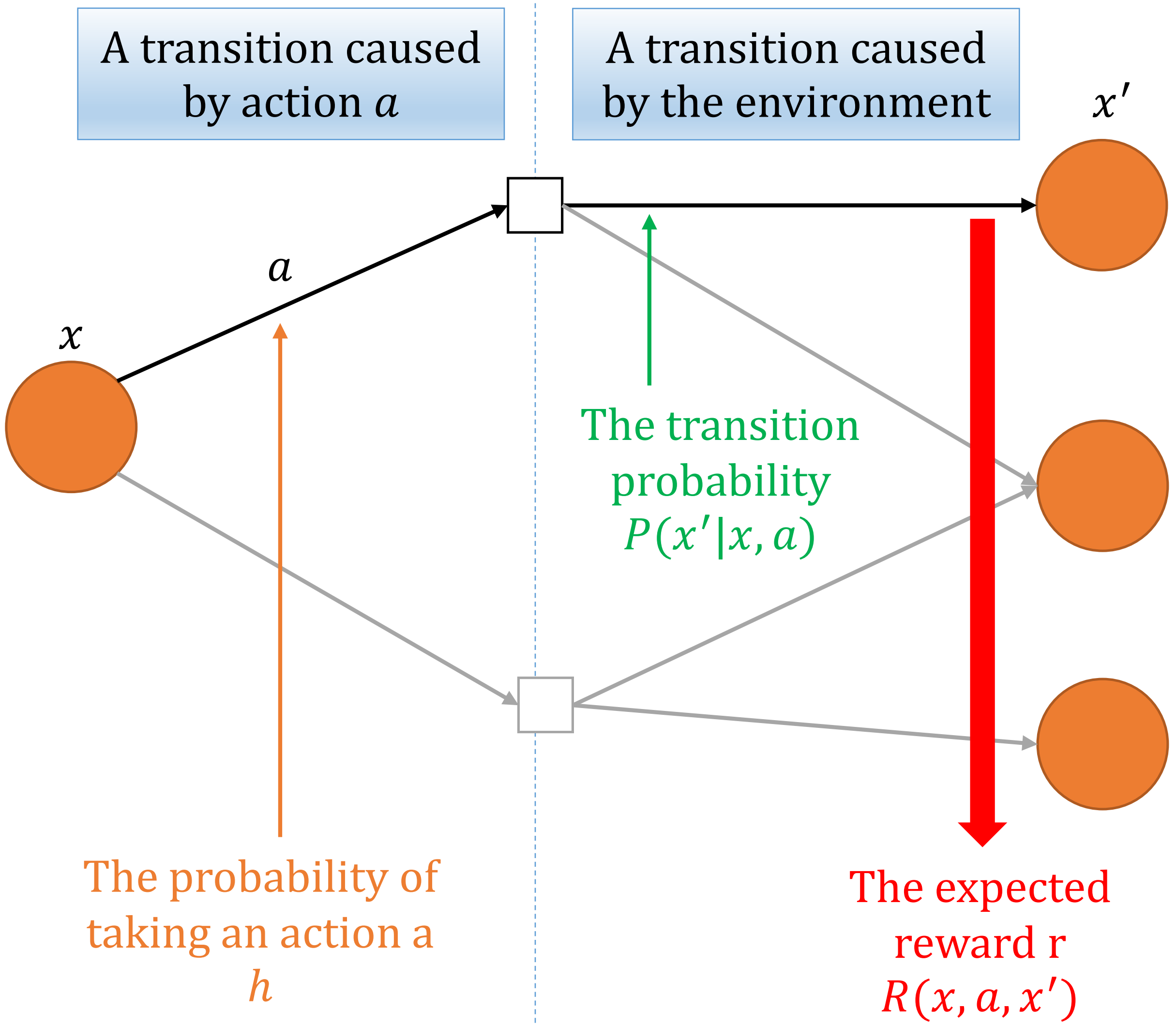
$R: X \times A \times X \rightarrow \mathbb{R}$ the reward function

An agent's policy h

h: the rule of an agent's selecting an action a

An agent want to choose the optimal policy h^* such that its behavior maximizes the discounted return at each time.

$$V_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k} \quad (\ \gamma \in [0,1)\ )$$



The Bellman optimality equation

If the policy h is stationary , the following equations are satisfied.

$Q: X \times A \to \mathbb{R}$ the action value function (Q-function)

(the expected discounted return of a state-action pair given the policy h)

The Bellman optimality equation

$$Q^*(x,a) = \sum_{x' \in X} P(x'|x,a)\left(R(x,a,x') + \gamma \max_a Q^*(x',a)\right) \qquad Q^*(x,a) = \max_h Q(x,a)$$

Q-learning

Q-*learning* is a learning algorithm of estimating $Q^*(x,a)$.

$$Q(x,a) \leftarrow Q(x,a) + \alpha\left[\underbrace{r + \gamma \max_{a\in A} Q(x',a)}_{\text{new Q-value}} - Q(x,a)\right] \quad (\alpha,\gamma \in (0,1]\text{:the learning rate})$$

Q-*learning* algorithm

Initialize $Q(x,a)$ for each Q-value and repeat the following steps.

1. Observe state x and decide $a \in \arg\max_{a} Q(x,a)$. ⟵ the policy h (stationary)
2. Acquire reward r and observe state transition to x'
3. Update $Q(x,a)$: $Q(x,a) \leftarrow Q(x,a) + \alpha\left[r + \gamma \max_{a\in A} Q(x',a) - Q(x,a)\right]$
4. t←t+1

Discrete Event Systems (DESs)

Reference
Tatsushi YAMASAKI and Toshimitsu USHIO ,Members,"Decentralized Supervisory Control of Discrete Event Systems Based on Reinforcement Learning",IEICE TRANS.FUNDAMENTALS,VOL.E88-A,NO.11 NOVEMBER 2005

A DES G : $< X, \Sigma, f, x_0 >$

X:a set of states

Σ:a set of events

$f: X \times \Sigma \rightarrow X$ a state transition function

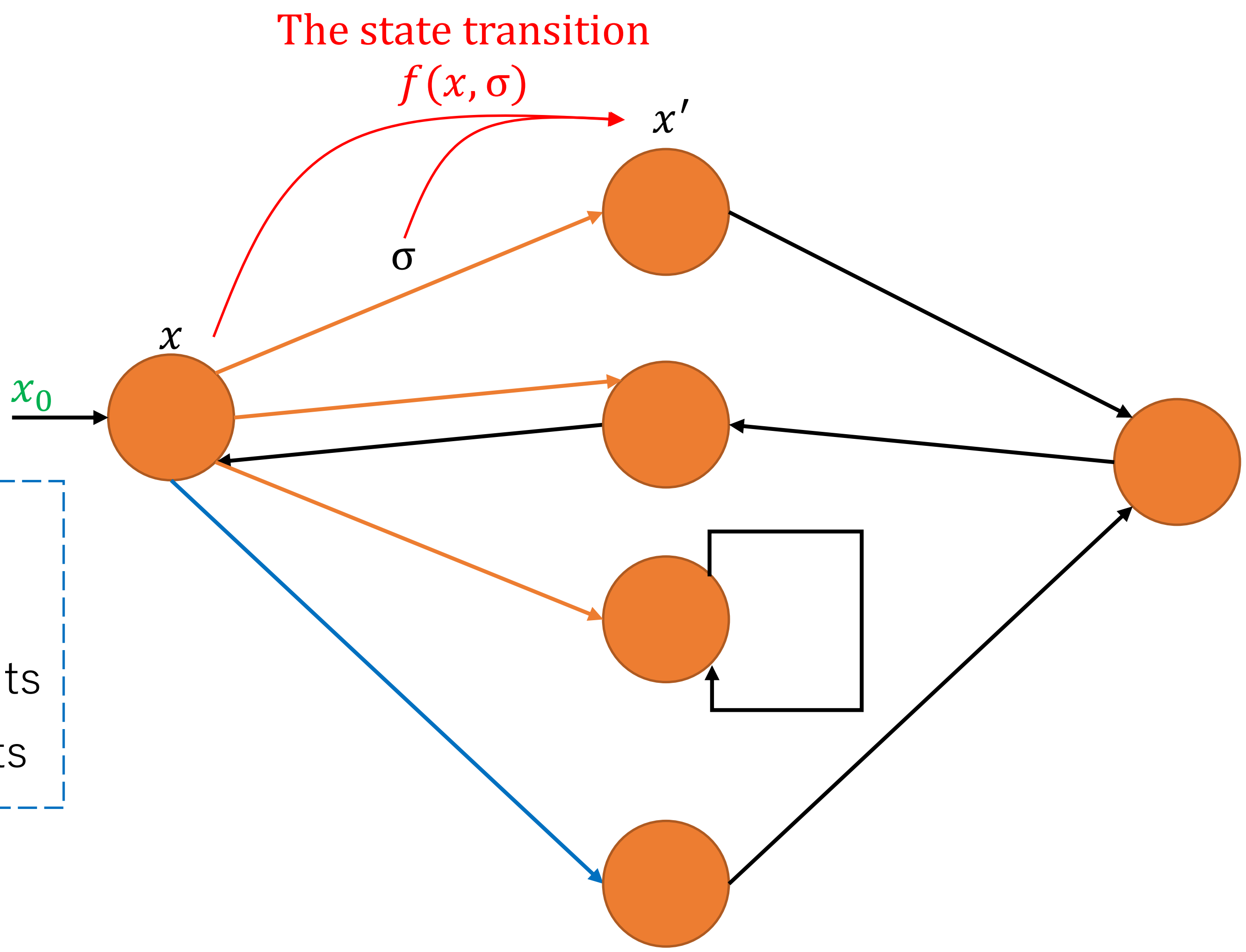
$x_0 \in X$:an initial state

$\Sigma^c \subseteq \Sigma$:a set of controllable events

$\Sigma^o \subseteq \Sigma$:a set of observable events

$\Sigma^{uc} = \Sigma - \Sigma^c$:a set of uncontrollable events

$\Sigma^{uo} = \Sigma - \Sigma^o$:a set of unobservable events

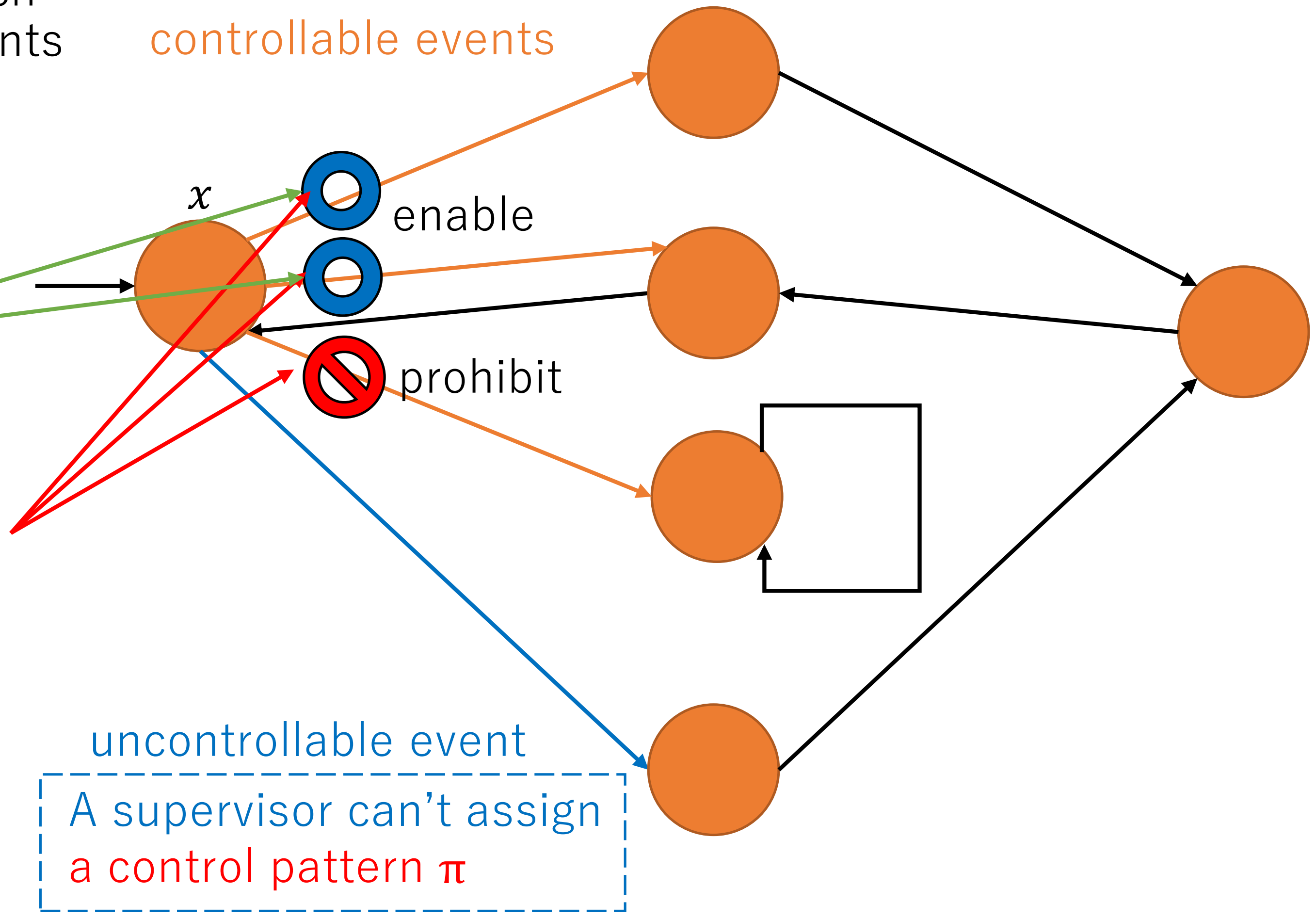


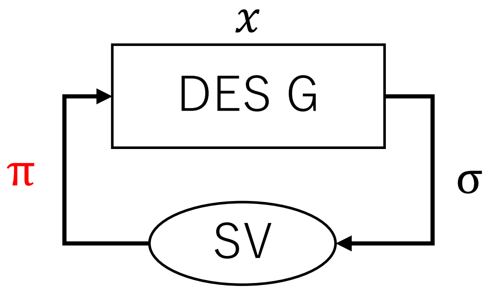
Supervisory control

Supervisors assign a control pattern π which enables or prohibits some controllable events so that a given specification is satisfied.

Supervisors don't cause a controllable event directly. (The environment cause a controllable event.)

a control pattern π at x





Decentralized supervisory control of DESs

$M_1^e: \Sigma \rightarrow \Sigma_i^o \cup \{\varepsilon\}$ the projection from σ in the DES G to σ_i for SV_i

a DES of Each local supervisor SV_i : $< S_i, \Sigma_i, g_i, x_0 >$

$S_i \subseteq 2^X$:the set of states

$\Sigma_i \subseteq \Sigma$:the set of events

$g_i: S_i \times \Sigma_i^o \rightarrow S_i$ the state transition function

$x_0 \in X$:the initial state of the DES G

$\Sigma_i^o \subseteq \Sigma^o$:the set of observable events for each SV_i

a MDP of SV_i : $< S_i, \Pi_i, P_i, R_i >$

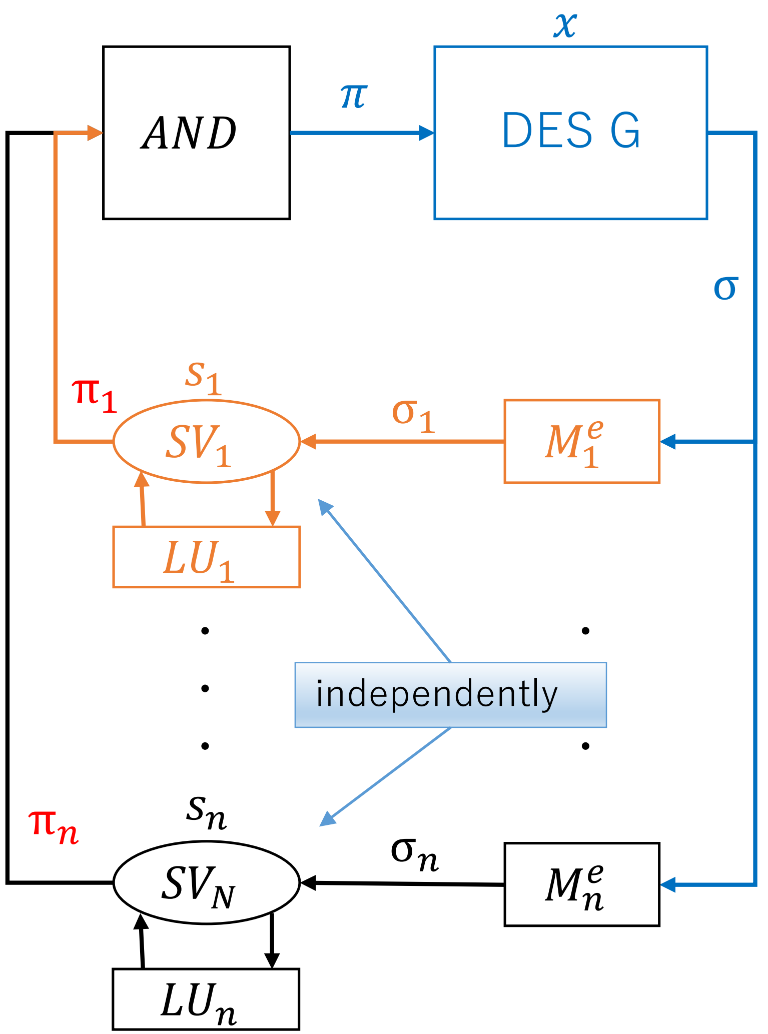
$S_i \subseteq 2^X$:the set of states of SV_i

Π_i:the set of control patterns at each state

$P_i: S_i \times \Pi_i \times S_i \rightarrow [0,1]$ the probability of the transition

$R_i: S_i \times \Pi_i \times S_i \rightarrow \mathbb{R}$ the expected reward

instead of an action



The system model based on Q-learning

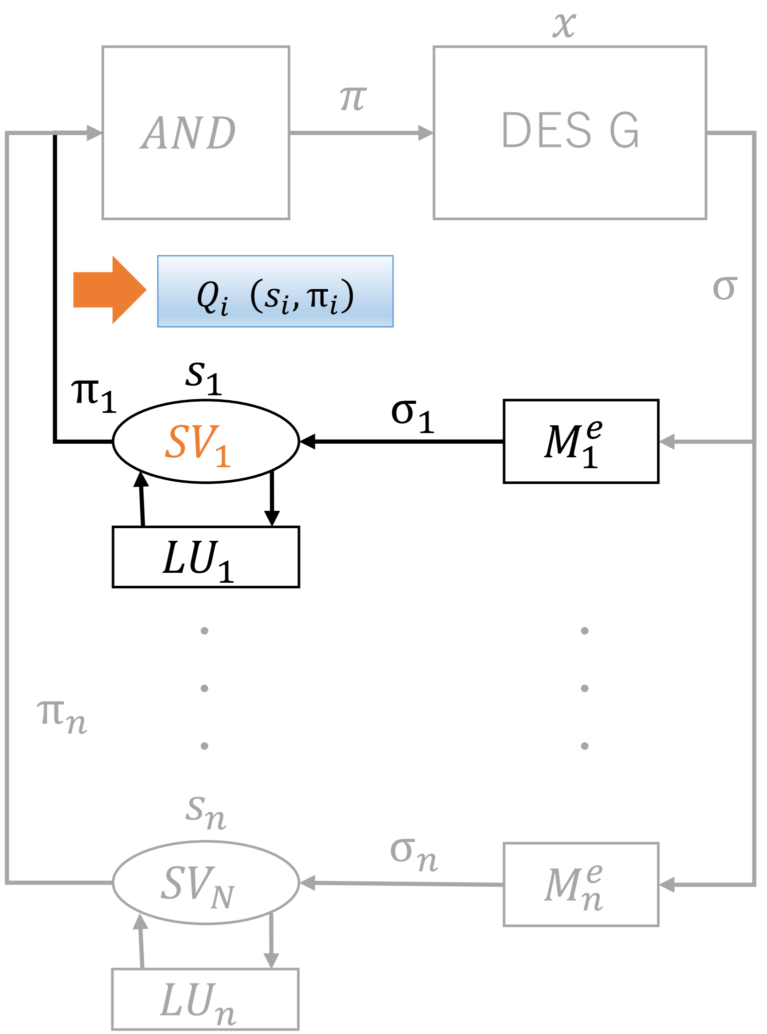
The Bellman optimal equation for each SV_i:

$Q_i : S_i \times \Pi_i \to \mathbb{R}$ the expected discounted return of a state-control pattern pair for SV_i

$$Q_i^*(s_i, \pi_i) = \sum_{s_i' \in S_i} P_i(s_i' | s_i, \pi_i) \left(R_i(s_i, \pi_i, s_i') + \gamma \max_{\pi_i' \in \Pi_i(s_i')} Q_i^*(s_i', \pi_i') \right)$$

$$Q_i^*(s_i, \pi_i) = \max_{\pi_i \in \Pi_i(s_i)} Q_i(s_i, \pi_i)$$

Q-learning



Two assumptions for the system (1/2)

1. For each SV_i , The following equation holds:

$$P_i(s_i'|s_i,\pi_i) = \sum_{\sigma_i^o \in \pi_i \cap \Sigma_i^o} P_i^1(s_i,\pi_i,\sigma_i^o) P_i^2(s_i,\sigma_i^o,s_i')$$

P_i^1:the probability of the occurrence of the observed event σ_i^o when SV_i selects π_i at s_i

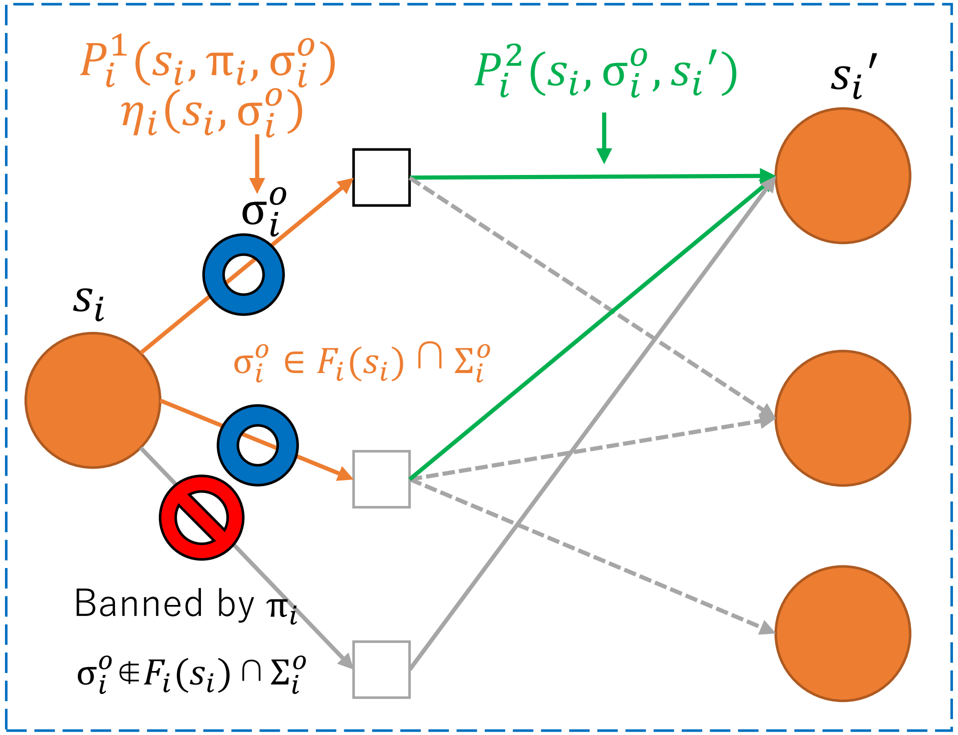
P_i^2:the probability of the transition from s_i to s_i' by the observed event σ_i^o

The DES G has a parameter $\eta_i(s_i,\sigma_i^o)$ which indicates a probability of the occurrence of the event σ_i^o at state s_i.

$$P_i^1(s_i,\pi_i,\sigma_i^o) = \frac{\eta_i(s_i,\sigma_i^o)}{\sum_{\sigma_i^o \in \pi_i \cap \Sigma_i^o} \eta_i(s_i,\sigma_i^o)}$$

$$\eta_i(s_i,\sigma_i^o) > 0 \qquad \sum_{\sigma_i^o \in F_i(s_i) \cap \Sigma_i^o} \eta_i(s_i,\sigma_i^o) = 1$$

SV_i selects π_i:



Two assumptions for the system (2/2)

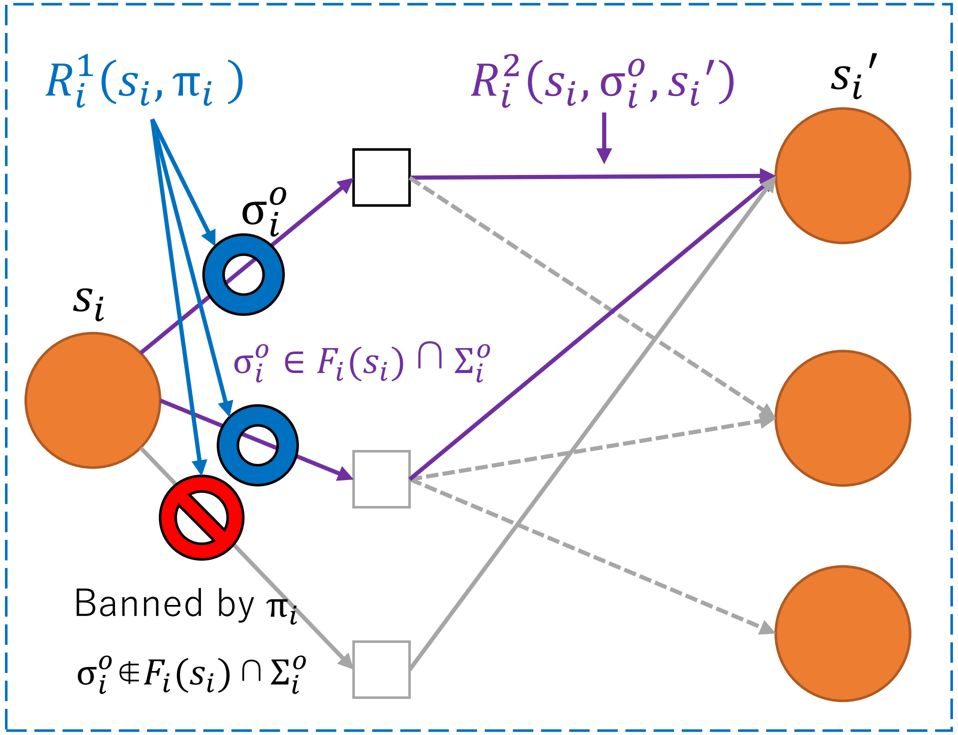
2. The reward $R_i(s_i, \pi_i, s_i')$ consists of two terms as follows:

$$R_i(s_i, \pi_i, s_i') = R_i^1(s_i, \pi_i) + R_i^2(s_i, \sigma_i^o, s_i')$$

R_i^1:the expected reward when SV_i selects π_i at s_i
→the cost to disable controllable events

R_i^2:the expected reward when SV_i observes an event σ_i^o and makes a transition from s_i to s_i'
→the costs by the occurrence of the event and evaluation about task

SV_i selects π_i:



Bellman optimal equation

By using the assumptions and Bellman optimal equation , the following equation is obtained:

$$Q_i^*(s_i, \pi_i) = R_i^1(s_i, \pi_i) + \sum_{\sigma_i^o \in \pi_i \cap \Sigma_i^o} \frac{\eta_i(s_i, \sigma_i^o)}{\sum_{\sigma_i^o \in \pi_i \cap \Sigma_i^o} \eta_i(s_i, \sigma_i^o)} T_i^*(s_i, \sigma_i^o)$$

$$T_i^*(s_i, \sigma_i^o) = \sum_{s_i' \in S_i} P_i^2(s_i, \sigma_i^o, s_i') \left(R_i^2(s_i, \sigma_i^o, s_i') + \gamma \max_{\pi_i' \in \Pi_i(s_i')} Q_i^*(s_i', \pi_i') \right)$$

similar to Q-learning

$T_i^*(s_i, \sigma_i^o)$ denotes a discounted expected total reward when SV_i observes σ_i^o at s_i and selects the control pattern which has the maximum value Q_i^* at the new states.

Bellman optimal equation

$$Q_i^*(s_i, \pi_i) = \sum_{s_i' \in S_i} P_i(s_i' | s_i, \pi_i) \left(R_i(s_i, \pi_i, s_i') + \gamma \max_{\pi_i' \in \Pi_i(s_i')} Q_i^*(s_i', \pi_i') \right)$$

Assumption 1.

$$P_i(s_i' | s_i, \pi_i) = \sum_{\sigma_i^o \in \pi_i \cap \Sigma_i^o} P_i^1(s_i, \pi_i, \sigma_i^o) P_i^2(s_i, \sigma_i^o, s_i')$$

$$P_i^1(s_i, \pi_i, \sigma_i^o) = \frac{\eta_i(s_i, \sigma_i^o)}{\sum_{\sigma_i^o \in \pi_i \cap \Sigma_i^o} \eta_i(s_i, \sigma_i^o)}$$

Assumption 2.

$$R(s_i, \pi_i, s_i') = R_i^1(s_i, \pi_i) + R_i^2(s_i, \sigma_i^o, s_i')$$

Bellman e.q. $\quad Q^*(x, a) = \sum_{x' \in X} P(x' | x, a) \left(R(x, a, x') + \gamma \max_a Q^*(x', a) \right)$

Q-learning Update $\quad Q(x, a) \leftarrow Q(x, a) + \alpha \left[r + \gamma \max_{a \in A} Q(x', a) - Q(x, a) \right]$

Formulation

Estimating $R_i^1(s_i,\pi_i)$, $\eta_i(s_i,\sigma_i^o)$ and $T_i\ (s_i,\sigma_i^o)$

$$T_i(s_i,\sigma_i^o) \leftarrow T_i(s_i,\sigma_i^o) + \alpha\left[r_i^2 + \gamma \max_{\pi_i{}' \in \Pi_i(s_i{}')} Q_i(s_i{}',\pi_i{}') - T_i(s_i,\sigma_i^o)\right]$$

$$R_i^1(s_i,\pi_i) \leftarrow R_i^1(s_i,\pi_i) + \beta\left[r_i^1 - R_i^1(s_i,\pi_i)\right]$$

for all $\sigma_i^o \in \pi_i \cap \Sigma_i^o$

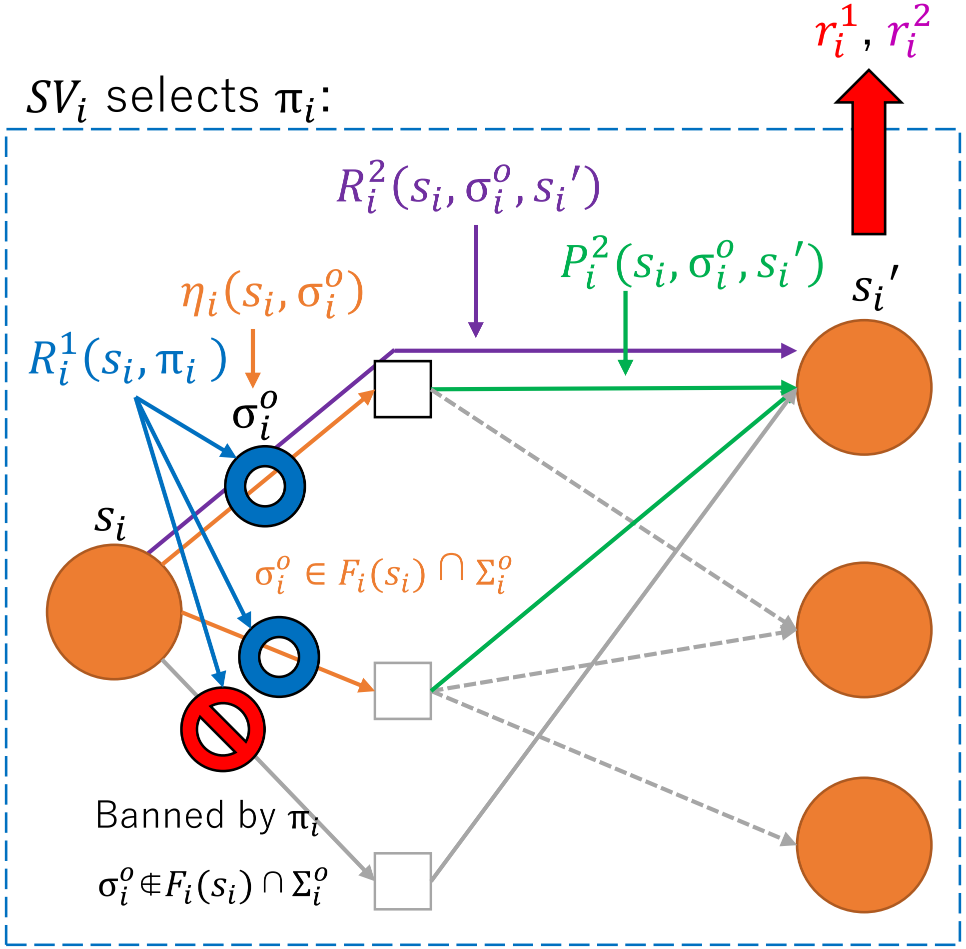
$$\eta_i(s_i,\sigma_i^o) \leftarrow \begin{cases} (1-\delta)\,\eta_i(s_i,\sigma_i^o{}') & \text{if} \quad \sigma_i^o{}' \neq \sigma_i^o \\ \eta_i(s_i,\sigma_i^o{}') + \delta\left[\sum_{\sigma_i^o \in \pi_i \cap \Sigma_i^o} \eta_i(s_i,\sigma_i^o) - \eta_i(s_i,\sigma_i^o{}')\right] & \text{if} \quad \sigma_i^o{}' = \sigma_i^o \end{cases}$$

Updating Q values

$\forall \pi_i{}' \in \Pi_i(s_i)\ s.t.\ \pi_i{}' \cap \pi_i \neq \emptyset$

$$Q_i\ (s_i,\pi_i{}') \leftarrow R_i^1(s_i,\pi_i{}') + \sum_{\sigma_i^o \in \pi_i \cap \Sigma_i^o} \frac{\eta_i(s_i,\sigma_i^o)}{\sum_{\sigma_i^o \in \pi_i \cap \Sigma_i^o} \eta_i(s_i,\sigma_i^o)} T_i\ (s_i,\sigma_i^o)$$

SV_i selects π_i:



The proposed algorithm

1. Initialize $R_i^1(s_i, \pi_i)$, $\eta_i(s_i, \sigma_i^o)$, $T_i\ (s_i, \sigma_i^o)$ and $Q_i\ (s_i, \pi_i)$ for all SV_i
2. Repeat until any s_i is a terminal state
   a. Initialize a state $s_i \leftarrow x_0$ for all SV_i
   b. Repeat for each SV_i
      i. Select a control pattern $\pi_i \in \Pi_i(s_i)$ based on the Q_i values by SV_i
      ii. Observe the occurrence of event $\sigma_i^o \in \Sigma_i^o$
      iii. Acquire rewards r_i^1 and r_i^2
      iv. Make a transition $s_i \rightarrow s_i'$ in SV_i
      v. Update $R_i^1(s_i, \pi_i)$, $\eta_i(s_i, \sigma_i^o)$ and $T_i\ (s_i, \sigma_i^o)$
      vi. Update $Q_i\ (s_i, \pi_i)$
      vii. $s_i \leftarrow s_i'$

Simulation : Setting (the cat and mouse problem)

a setting of states

A mouse can move in room1 , room2 and room3.

A cat can move in room3 , room4 and room5.

a setting of SV_1

SV_1 can observe the occurrences of the event

$\sigma_i^o \in \Sigma_1^o = \{m1, m2, m3, c2, c3\}$ in the room1 , room2 and room3.

SV_1 can control $m1, m2$ and $m3$

a setting of SV_2

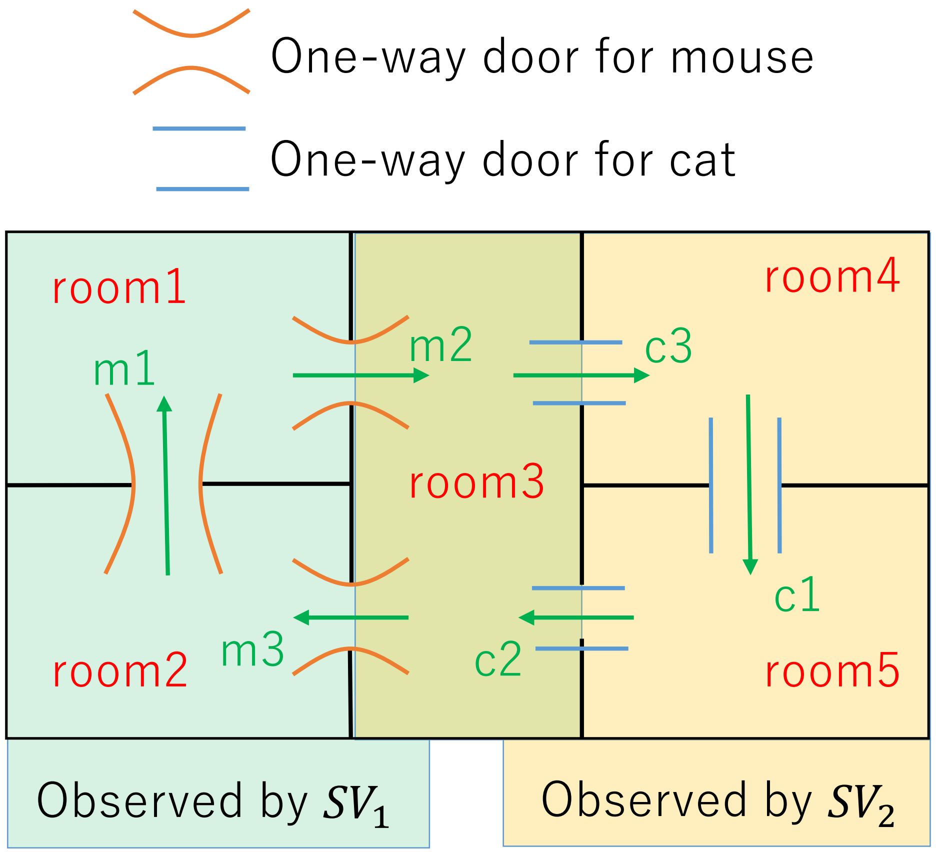
SV_2 can observe the occurrences of the event

$\sigma_i^o \in \Sigma_2^o = \{c1, c2, c3, m2, m3\}$ in the room3 , room4 and room5.

SV_2 can control $c1, c2$ and $c3$

This problem's goal

controlling doors so as not to encounter a cat and a mouse in the same room simultaneously



Simulation : Setting (the cat and mouse problem)

the initial state

x_0=(room2 , room4)
mouse cat

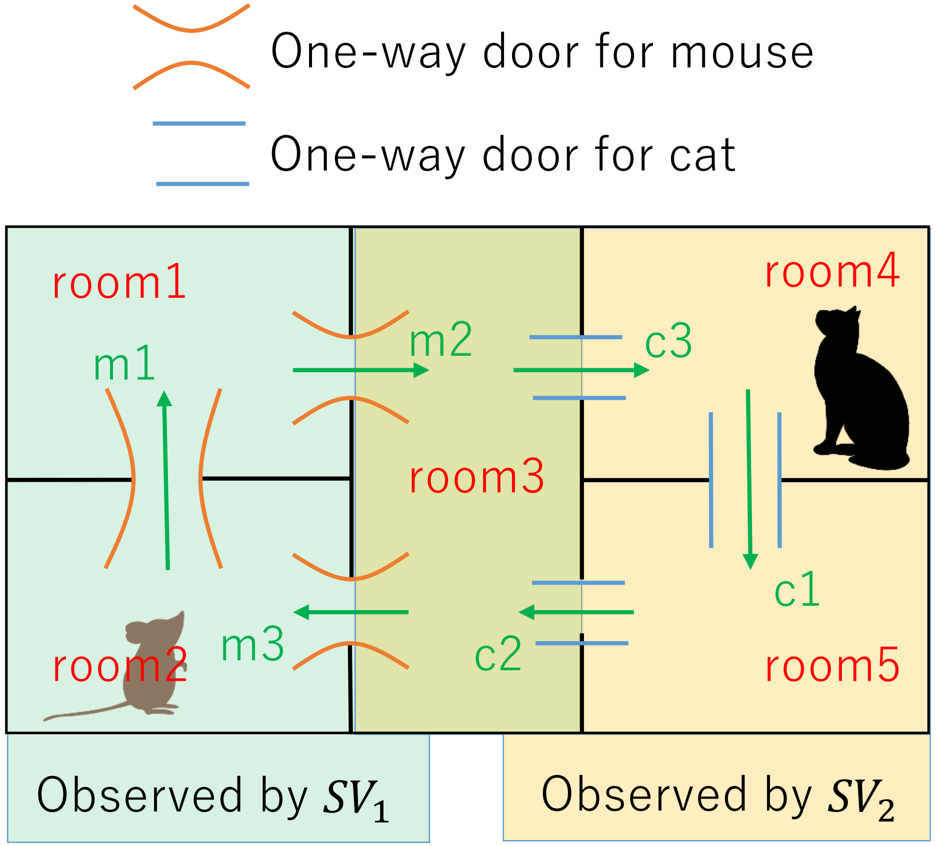
a reward r_i^1 and r_i^2

$r_i^1 = -2 \times$(the number of doors closed by SV_i)

$$r_i^2 = \begin{cases} 1 & \text{if } SV_i \text{ observes a cat and a mouse entering a new room} \\ -100 & \text{if } SV_i \text{ observes a cat and a mouse in room3} \\ 0 & \text{otherwise} \end{cases}$$

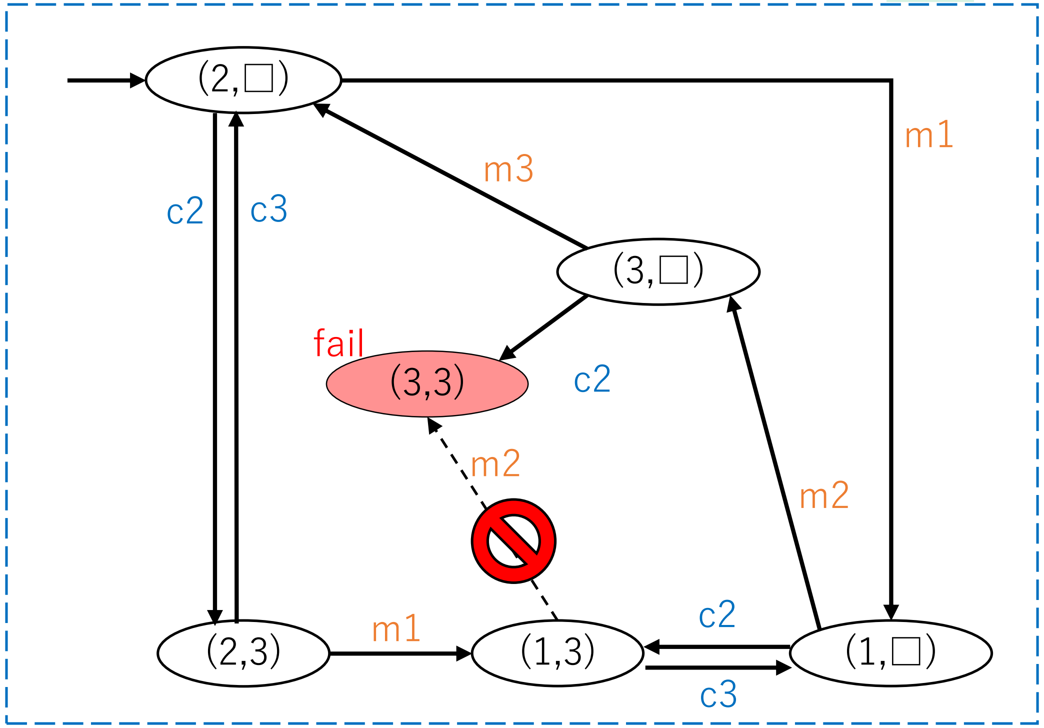
observation noise (the normal distribution)
μ=(true value) σ=0.1

SV_i prefers to leave doors open
if the encounter does not occur.



Simulation : Result

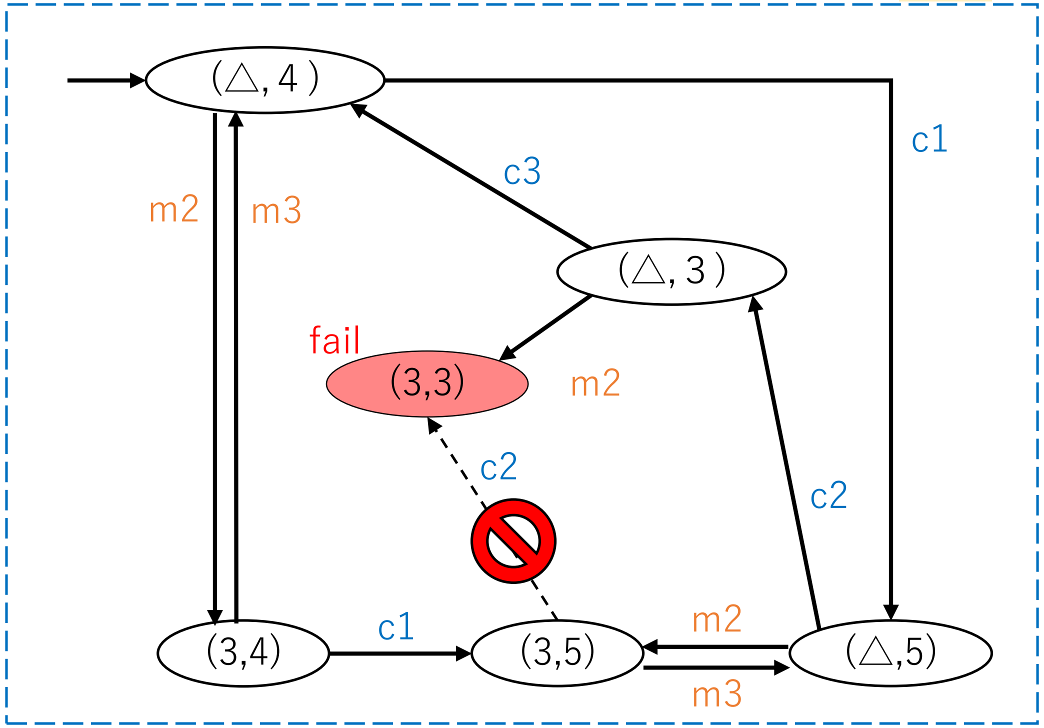
The transition diagram of the learned SV_1



mi : controllable ci : uncontrollable

□ : the unobservable state (4 or 5)

The transition diagram of the learned SV_2



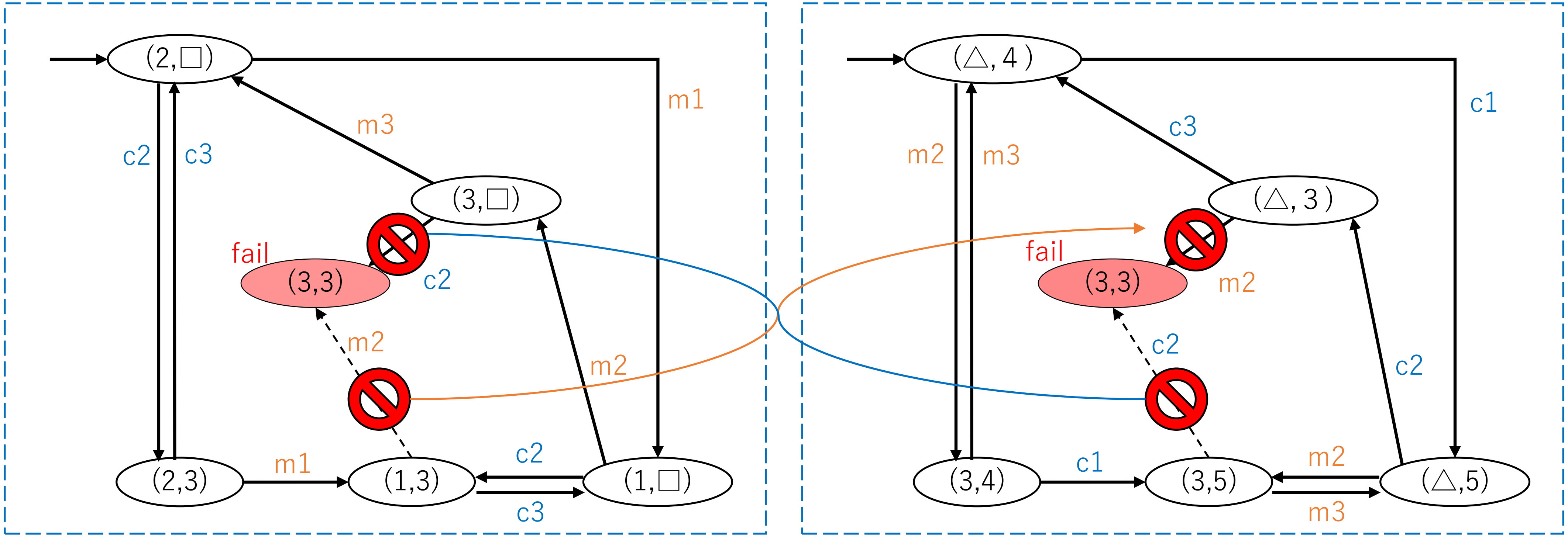
mi : uncontrollable ci : controllable

△ : the unobservable state (1 or 2)

Simulation : Result

The transition diagram of the learned SV_1

The transition diagram of the learned SV_2



mi : controllable ci : uncontrollable

□ : the unobservable state (4 or 5)

mi : uncontrollable ci : controllable

△ : the unobservable state (1 or 2)

Future Work

- We propose a decentralized supervisory control method based on RL such that the control systems satisfy a LTL specifications.
- Simulation