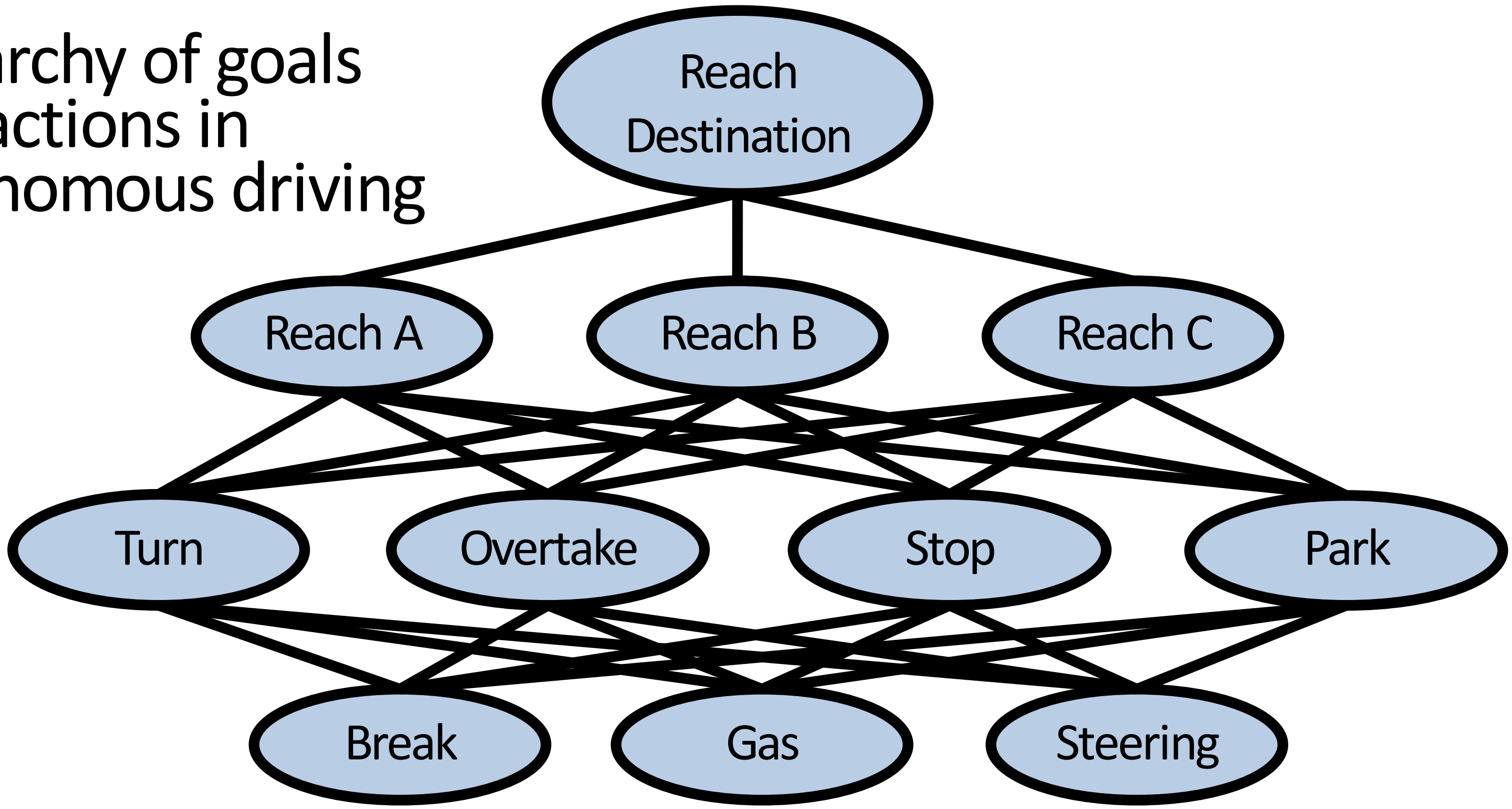
CS885 Reinforcement Learning
Lecture 15c: June 20, 2018

Semi-Markov Decision Processes

[Put] Sec. 11.1-11.3

Hierarchical RL

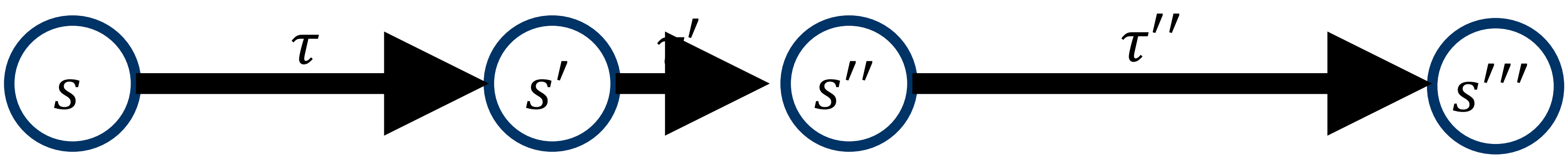
- Hierarchy of goals and actions in autonomous driving



- Theory: Semi-Markov Decision Processes

Semi-Markov Process

- Definition
  - Set of States: S
  - Transition dynamics: $\Pr(s', \tau|s) = \Pr(s'|s)\Pr(\tau|s)$ where τ indicates the time to transition
- Semi-Markovian:
  - Next state depends only on current state
  - Time spent in each state varies



Semi-Markov Decision Process

- Definition
  - Set of states: S
  - Set of actions: A
  - Transition model: $\Pr(s', \tau | s, a)$
  - Reward model: $R(s, a) = E[r | s, a]$
  - Discount factor: $0 \leq \gamma \leq 1$
    - discounted: $\gamma < 1$ undiscounted: $\gamma = 1$
  - Horizon (i.e., # of time steps): h
    - Finite horizon: $h \in \mathbb{N}$ infinite horizon: $h = \infty$

- Goal: find optimal policy

Example from Queuing Theory

- Consider a retail store with two queues:
  - Customer service queue
  - Cashier queue
- Semi-Markov decision process
  - State: $s = (q_1, q_2)$ where $q_i =$ # of customers in queue i
  - Action: $a \in \{1,2\}$ (i.e., serve customer in queue 1 or 2)
  - Transition model: distribution over arrival and service times for customers in each queue.
  - Reward model: expected revenue of each serviced customer – expected cost associated with waiting times
  - Discount factor: $0 \leq \gamma < 1$
  - Horizon (i.e., # of time steps): $h = \infty$

Value Function and Policy

- Objective: $V^{\pi}(s) = \sum_i \gamma^{t_i} E\left[R\left(s_{t_i}, \pi(s_{t_i})\right)\right]$
  - Where $t_i = \tau_0 + \tau_1 + \cdots + \tau_i$
  - Optimal policy: π^* such that $V^{\pi^*}(s) \geq V^{\pi}(s)\ \forall s, \pi$

- Bellman's equation:

$$V^*(s) = \max_a R(s,a) + \sum_{s',\tau} \Pr(s', \tau | s, a)\, \gamma^{\tau} V^*(s')$$

- Q-learning update:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma^{\tau} \max_{a'} Q(s',a') - Q(s,a)\right]$$

Option Framework

- Semi-Markov decision process where actions are options (temporally extended sub-policies)

- Let a be an option with sub-policy π and terminal states S_{end}

$$\forall s_{t+\tau} \in S_{end}: \Pr(s_{t+\tau}, \tau | s_t, a) = \sum_{s_{t+1:t+\tau-1} \notin S_{end}} \prod_{i=1}^{\tau-1} \Pr(s_{t+i} | s_{t+i-1}, \pi(s_{t+i-1}))$$

$$R(s_t, a, s_{t+\tau}, \tau) = R\big(s_t, \pi(s_t)\big) + \gamma \sum_{s_{t+1}} \Pr(s_{t+1} | s_t, \pi(s_t)) \Big[R\big(s_{t+1}, \pi(s_{t+1})\big) + \cdots \gamma \sum_{s_{t+\tau}} \Pr(s_{t+\tau} | s_{t+\tau-1}, \pi(s_{t+\tau-1})) \big[R\big(s_{t+\tau}, \pi(s_{t+\tau})\big)\big] \ldots\Big]$$

Option Framework

- Bellman's equation:

$$V^*(s) = \max_a \sum_{s',\tau} \Pr(s',\tau|s,a)\left[R(s,a,s',\tau) + \gamma^\tau V^*(s')\right]$$

- Q-learning update:

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left[G_\tau + \gamma^\tau \max_{a'} Q(s',a') - Q(s,a)\right]$$

where $G_\tau = \sum_{i=0}^{\tau} \gamma^i r_i$