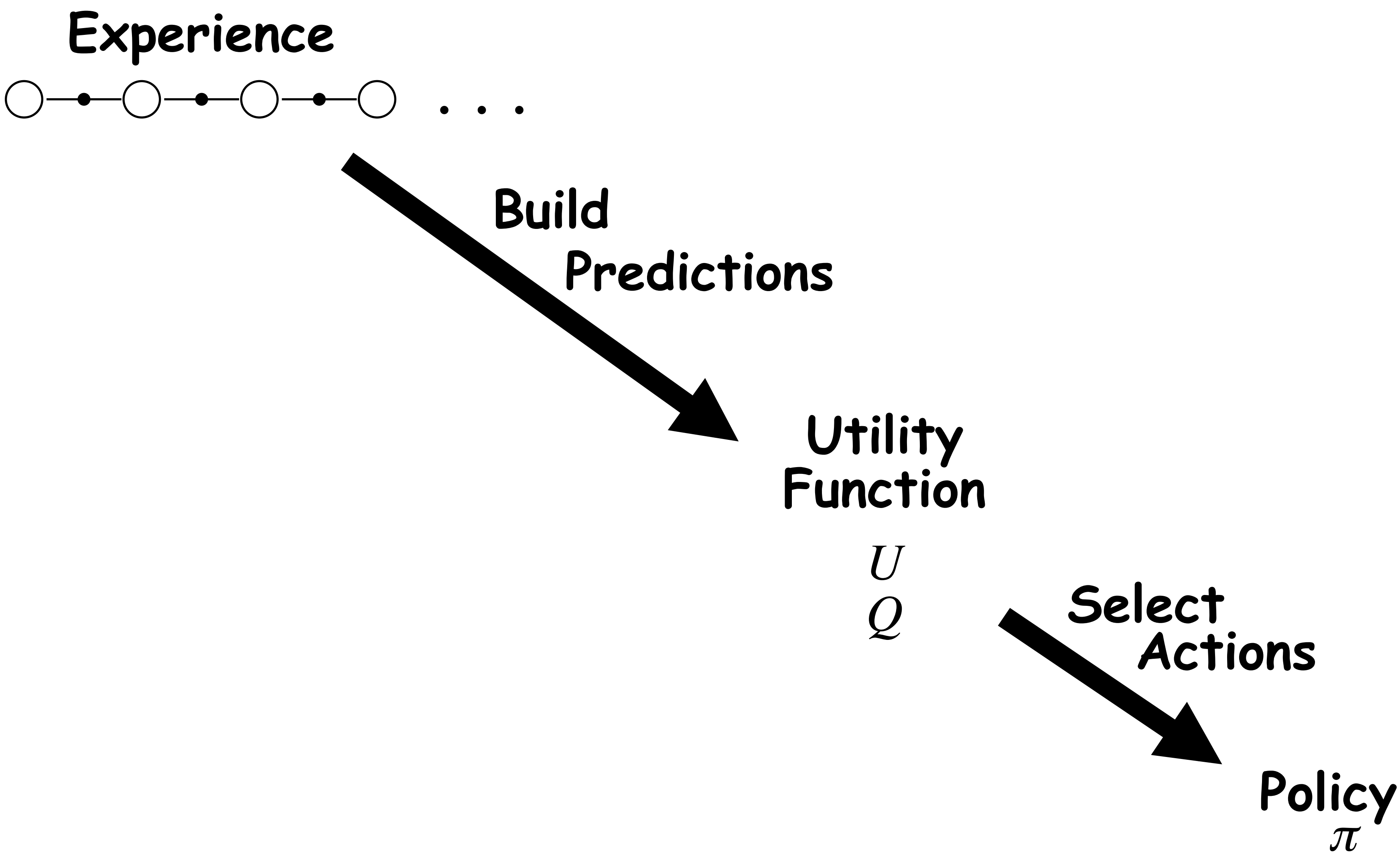
Reinforcement Learning for HW 5

CMPSCI 383
Nov 29, 2011

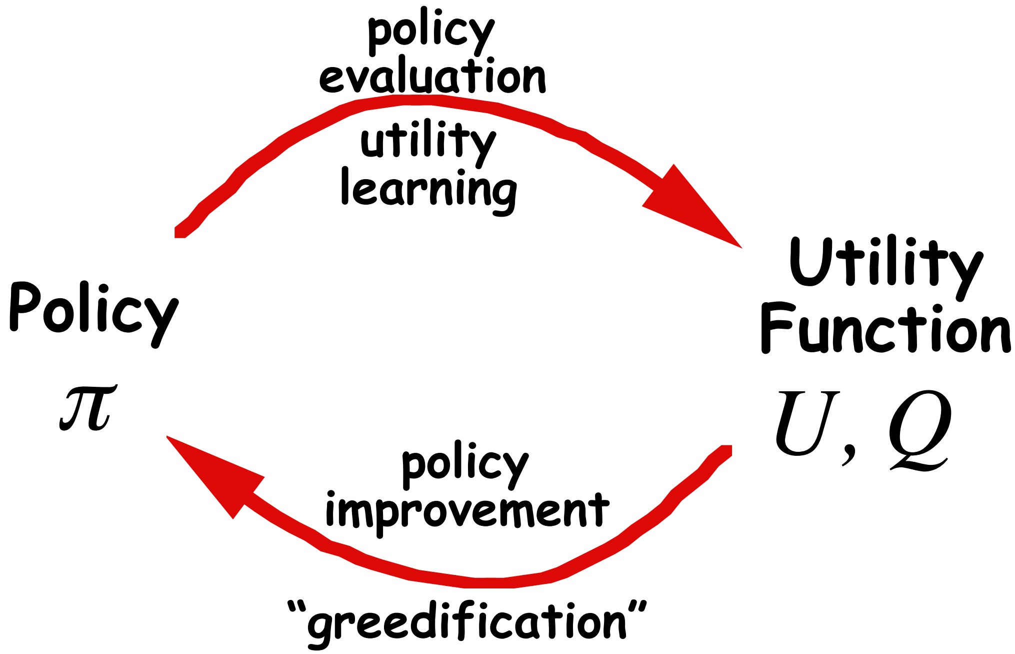
Today's lecture

- Active agents
- The exploration/exploitation dilemma
- Q-Learning

Active RL Agents



Interaction of policy and utility



What is Q?   Action-value function

$$Q(s,a) =$$

Utility of doing action *a* in state *s*
i.e.: Total amount of reward expected over the future if you do action *a* in state *s* and thereafter select optimal actions.

The utility of a state is the utility of doing the best action from that state:

$$U(s) = \max_a Q(s,a)$$

Learning an action-value function

- Q-Learning directly assigns a Q-value, Q(s,a), to each [state,action] pair.
- Don't need to learn transition probabilities to decide on best action:

$$\pi^*(s) = \arg\max_a Q(s,a)$$

Bellman Equation for Q functions

Recall Bellman Equation for *U*:

$$U(s) = R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s' \mid s, a) U(s')$$

$$Q(s,a) = R(s) + \gamma \sum_{s'} P(s' \mid s,a) \max_{a'} Q(s',a')$$

Q-Learning

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left(R(s) + \gamma \max_{a'}(s',a') - Q(s,a) \right)$$

SARSA

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left(R(s) + \gamma\, Q(s',a') - Q(s,a) \right)$$

Sarsa

Initialize $Q(s, a)$ arbitrarily
Repeat (for each episode):
    Initialize s
    Choose a from s using policy derived from Q (e.g., ϵ-greedy)
    Repeat (for each step of episode):
        Take action a, observe r, s'
        Choose a' from s' using policy derived from Q (e.g., ϵ-greedy)
        $Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma Q(s',a') - Q(s,a)\right]$
        $s \leftarrow s'$; $a \leftarrow a'$;
    until s is terminal

What's the best exploration policy?

Assume you've learned a utility function,
How do you select actions?

Greedy Action Selection:

Always select the action that looks best:

$$\pi(s) = \arg\max_{a} Q(s,a)$$

ε-Greedy Action Selection:

Be greedy most of the time

Occasionally take a random action

Other Methods:

Boltzmann distribution, Keep track of confidence intervals, etc.

Exploitation vs Exploration!

ε-Greedy Action Selection

- Greedy action selection:

$$a_t = a_t^* = \arg\max_a Q_t(a)$$

- ε-Greedy:

$$a_t = \begin{cases} a_t^* & \text{with probability } 1-\varepsilon \\ \text{random action} & \text{with probability } \varepsilon \end{cases}$$

. . . the simplest way to try to balance exploration and exploitation