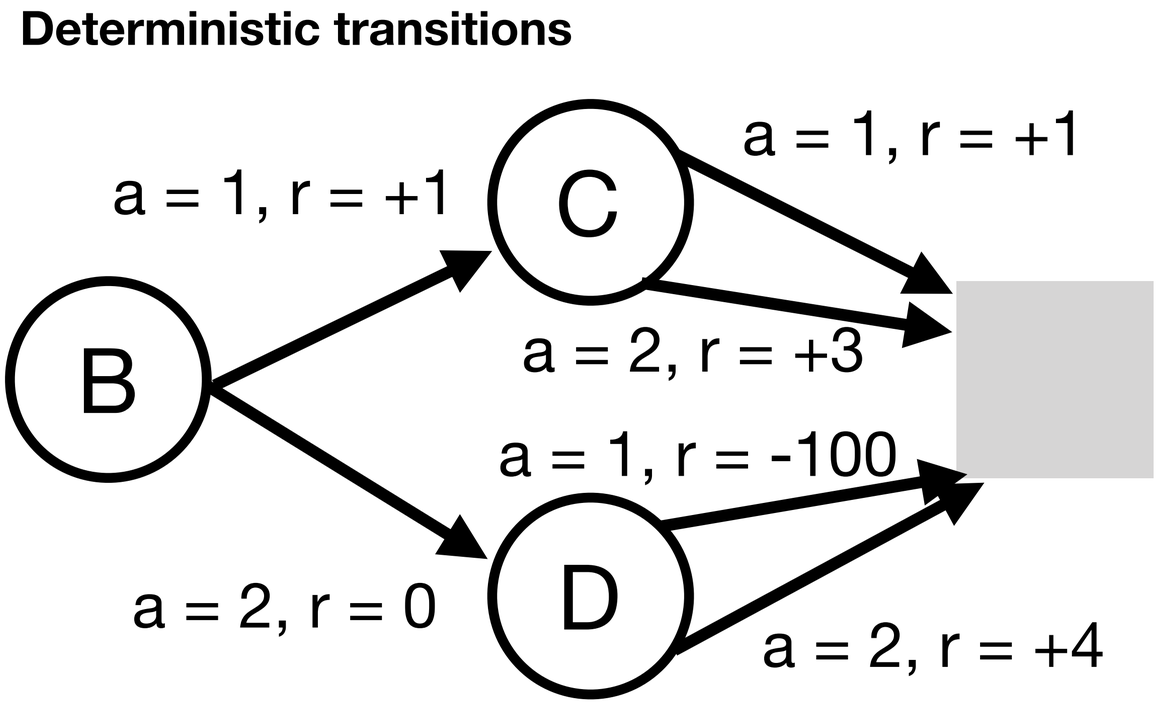
Course 2, Module 3
Temporal Difference Learning Methods for Control

CMPUT 397
Fall 2019

1. Consider the following MDP, with three states B, C and D ($\mathcal{S} = \{B, C, D\}$), and 2 actions ($\mathcal{A} = \{1, 2\}$), with $\gamma = 1.0$. Assume the action values are initialized $Q(s, a) = 0 \; \forall \; s \in \mathcal{S}$ and $a \in \mathcal{A}$. The agent takes actions according to an ϵ-greedy with $\epsilon = 0.1$.

(a) What is the optimal policy for this MDP and what are the action-values corresponding to the optimal policy: $q^*(s, a)$?

(b) Imagine the agent experienced a single episode, and the following experience: $S_0 = B, A_0 = 2, R_1 = 0, S_1 = D, A_1 = 2, R_2 = 4$. What are the Sarsa updates during this episode? Start with state B, and perform the Sarsa update, then update the value of state D.

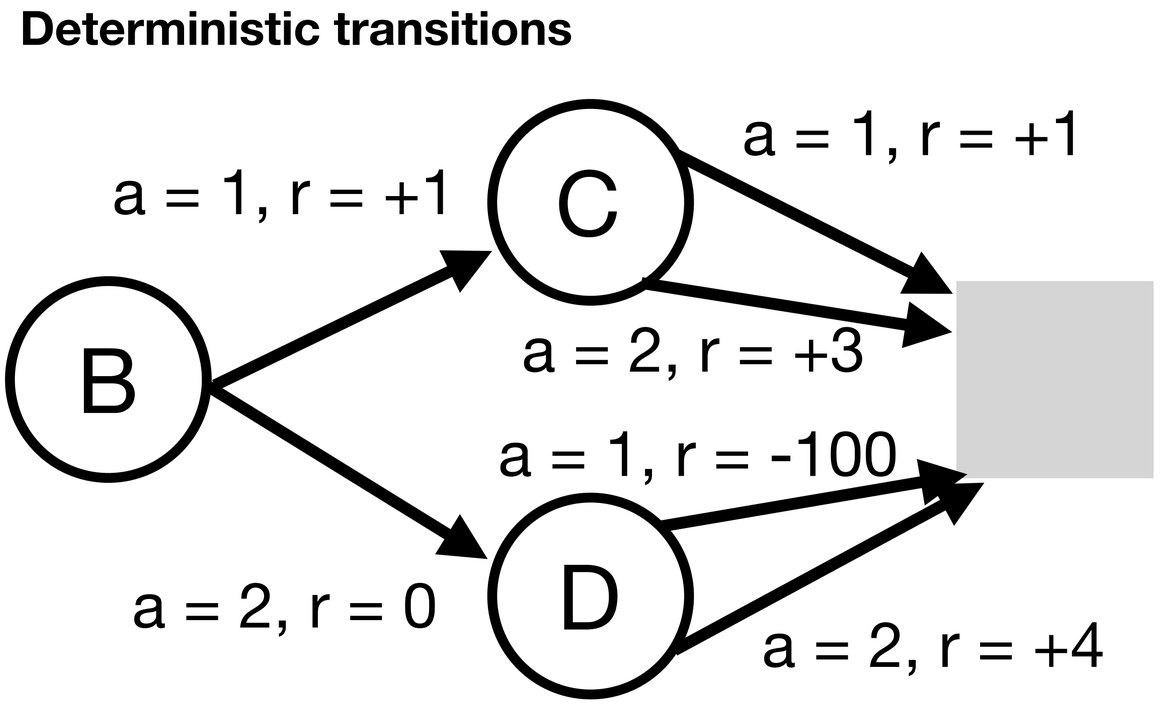


(b) Imagine the agent experienced a single episode, and the following experience: $S_0 = B, A_0 = 2, R_1 = 0, S_1 = D, A_1 = 2, R_2 = 4$. What are the Sarsa updates during this episode? Start with state B, and perform the Sarsa update, then update the value of state D.

(c) Using the sample episode above, compute the updates Q-learning would make. Again start with state B, and then state D.

(d) Let's consider one more episode: $S_0 = B, A_0 = 2, R_1 = 0, S_1 = D, A_1 = 1, R_2 = -100$. What would the Sarsa updates be? And what would the Q-learning updates be?

(e) What policy does Q-learning converge to? What policy does Sarsa converge to?



Understanding Expectation and Variance

- We talk about the differences in the variance of updates
- **First**: Let's compare the variance of the target for Sarsa and Expected Sarsa
- **Second**: Let's go back to the question we did before, looking at the variance of the TD update and Monte Carlo update

5. In this question we compare the variance of the target for Sarsa and Expected Sarsa. Recall the update for Sarsa is

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t) \right]$$

and for Expected Sarsa is

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \sum_{a' \in \mathcal{A}} \pi(a'|S_{t+1}) Q(S_{t+1}, a') - Q(S_t, A_t) \right].$$

(a) Start by comparing the part of the update that is different: $Q(S_{t+1}, A_{t+1})$ compared to $\sum_{a' \in \mathcal{A}} \pi(a'|S_{t+1}) Q(S_{t+1}, a')$. Write down the variance for these two terms, given $S_{t+1} = s'$.

$$\text{Var}(Q(s', A_{t+1})) \qquad \text{and} \qquad \text{Var}\left(\sum_{a' \in \mathcal{A}} \pi(a'|s') Q(s', a') \right)$$

Conclude that the variance is zero for Expected Sarsa, but likely non-zero for Sarsa. Notice that the only random variable is A_{t+1}, which is the action selected according to the target policy π with distribution $\pi(\cdot|S_{t+1})$.

2. In Monte Carlo control, we required that every state-action pair be visited infinitely often. One way this can be guaranteed is by using exploring starts. Can we use exploring starts for Sarsa? Further, we have talked about using Sarsa with an ϵ-greedy policy. Can we use Monte Carlo with an ϵ-greedy policy? Does this ensure sufficient exploration?