Course 2, Module 3
Temporal Difference Learning Methods for Control

CMPUT 397
Fall 2019

Any questions about course admin?

- Link for questions:
  - **http://www.tricider.com/brainstorming/35B8Mn3NZ5B**

Q&A / Clarifications of Course 2, Module 3
TD Control

Preciseness is important

- When submitting **discussion questions** ...
- When answering **quiz questions** ...
- When writing the **midterm and exam** ...
    - using the correct terminology & proper spelling and grammar matters!
- There is a lot of terminology in this course....**but not nearly as much** as in Biology or Neuroscience
- It's not ok to say *mammal* when we mean *reptile*. It's not ok so write *sorting algorithm* when we mean finding the *max*

We will test you on Preciseness

- It's clear from many of discussion question, that some are not taking care to:
  - make their sentences grammatically correct, or use the correct terminology
- **11 submission received a grade of zero.** If I cannot understand what you are asking, or if you asked a question directly answered in the textbook reading or video. **Zero**
- Preciseness is important in Computing Science

Mini-test

Q1) A problem without terminations (not episodic) is called a _______ problem?

a) continuous

b) policy evaluation

c) control

d) continuing

Clarifications

- Can we use Q-learning or Sarsa in an environment without **terminations**? >> yep
- Would we prefer Sarsa to Expected Sarsa if **|A| was large**? >> same
- Does the extra computation in Expected Sarsa really matter? >> not really
- Why does Q-learning ***ignore*** exploratory actions? >> does it?
- How does Q-learning ***update*** its behavior policy? >> through Q
- Would TD-control methods be useful on other problems **outside RL**? >> yep

Clarifications

- In the cliff world with Q-learning, why not just *switch to the greedy policy* after a while? >> good idea
- Can Expected Sarsa *switch* between on and off policy modes? How would that work? >> yep
- Could we use the ***variance*** instead of the ***expected value*** inside Expected Sarsa? Q(S,A) = Q(S,A) + alpha[R + gamma*VAR[Q(S', .)] - Q(S,A)]
- Why do we use small step-size values (small alpha)? >> variance!
- Why does Sarsa do poorly / diverge with large alpha? >> variance!

Clarifications

- What special considerations would we have to consider for using TD control algorithms in **non-stationary domains**?

- What target policy could Expected Sarsa use that would be:
  (a) **not the greedy policy**, and
  (b) not the same as the behavior policy?

- Why is Q-learning more popular than Expected Sarsa? >> new vs old-but-good

Longer Clarification

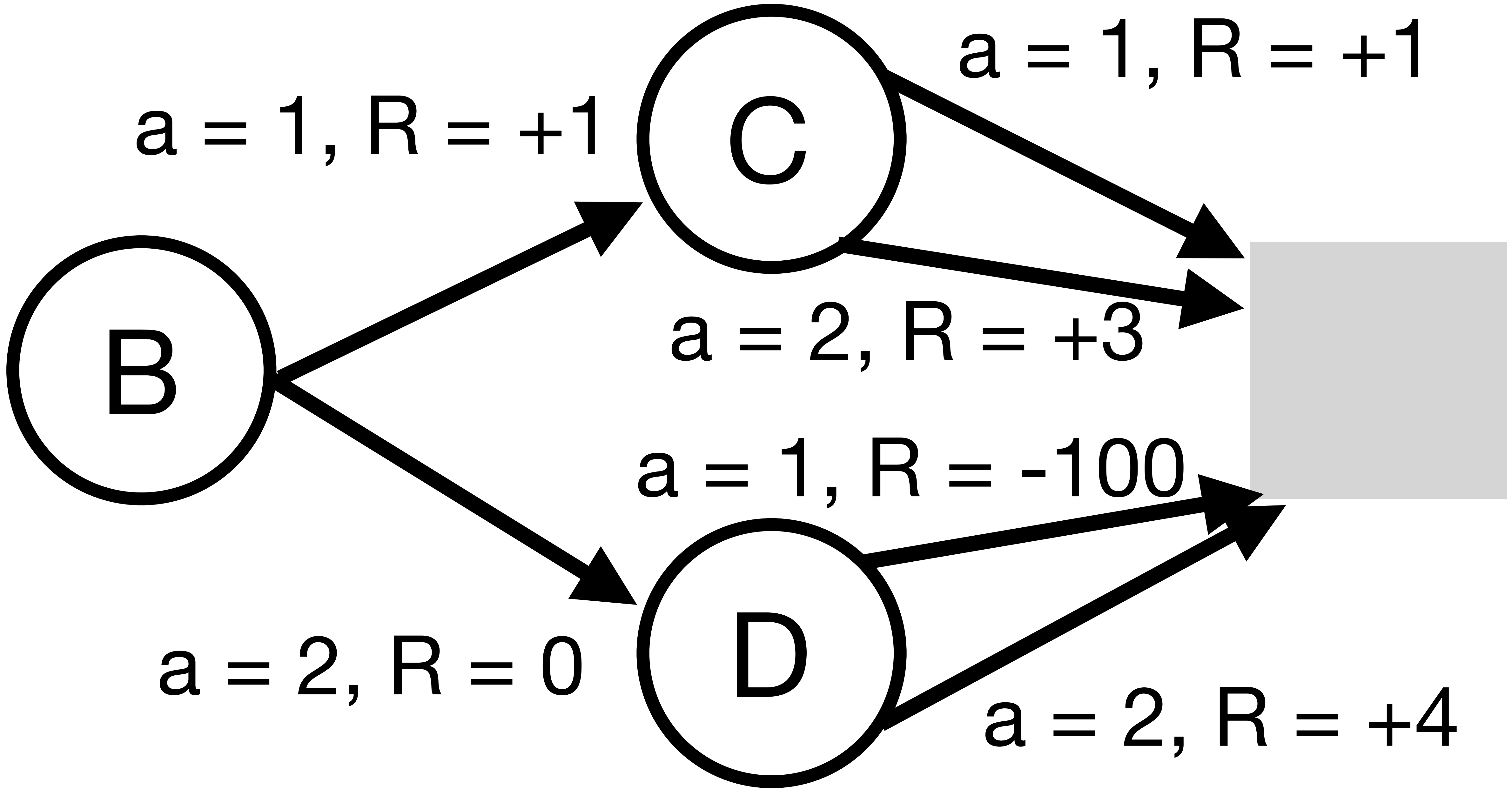
- How could **Monte Carlo control get stuck** and never succeed in the Windy gridworld?
- Why does the value of alpha have a *bigger impact on Sarsa* compared with Expected Sarsa?
- **Why does Sarsa learn to take the longer path? >> lets see why in an exercise**

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.

Deterministic transitions



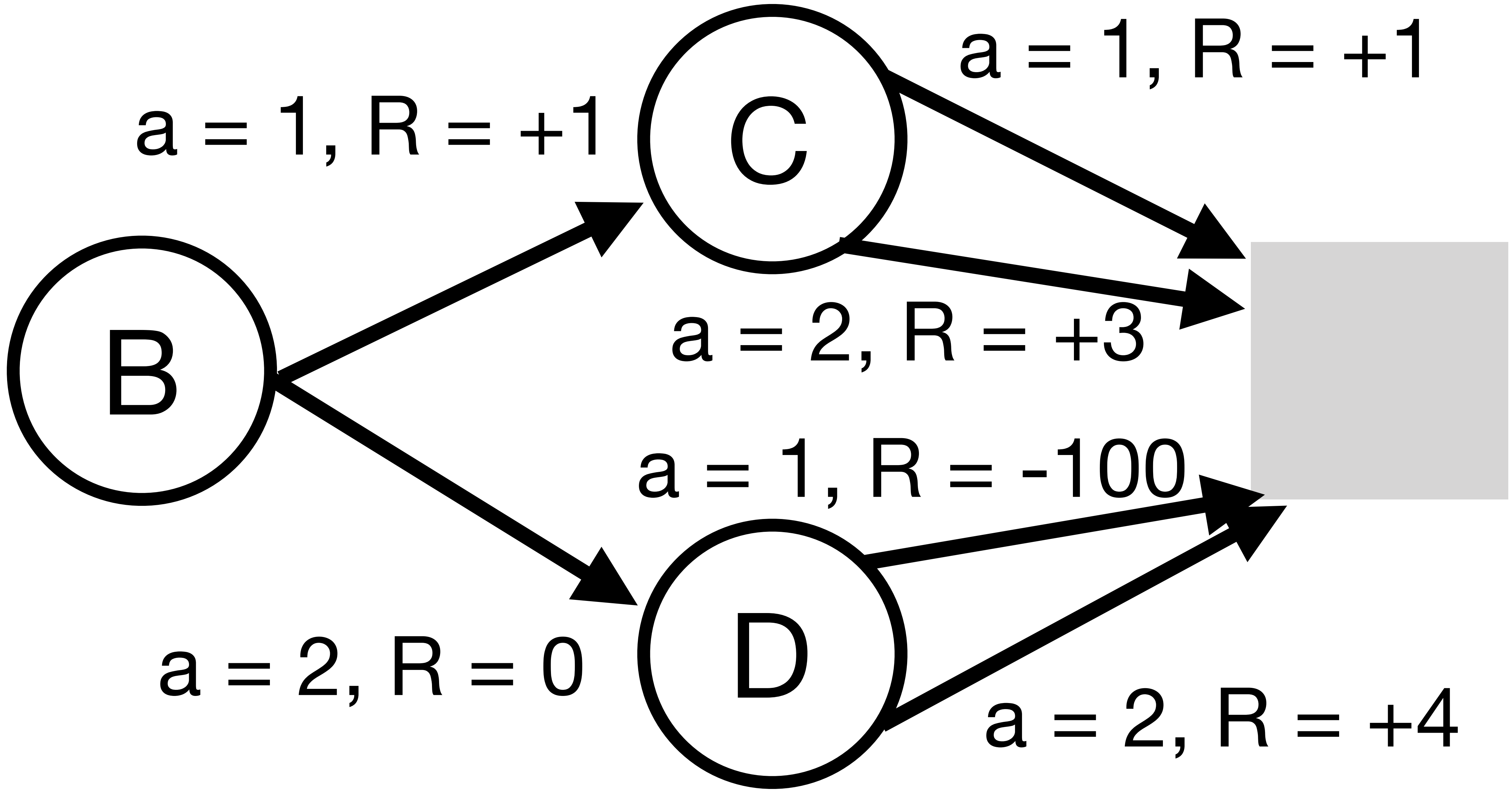
(b) Imagine the agent experienced a single episode, and the following experience: $S_0 = B, A_0 = 2, R_1 = 0, S_1 = D, A_1 = 2, R_2 = 4$. What are the Sarsa updates during this episode? Start with state B, and perform the Sarsa update, then update the value of state D.

(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?

Deterministic transitions


a = 1, R = +1
C
a = 1, R = +1
B
a = 2, R = +3
a = 1, R = -100
a = 2, R = 0
D
a = 2, R = +4