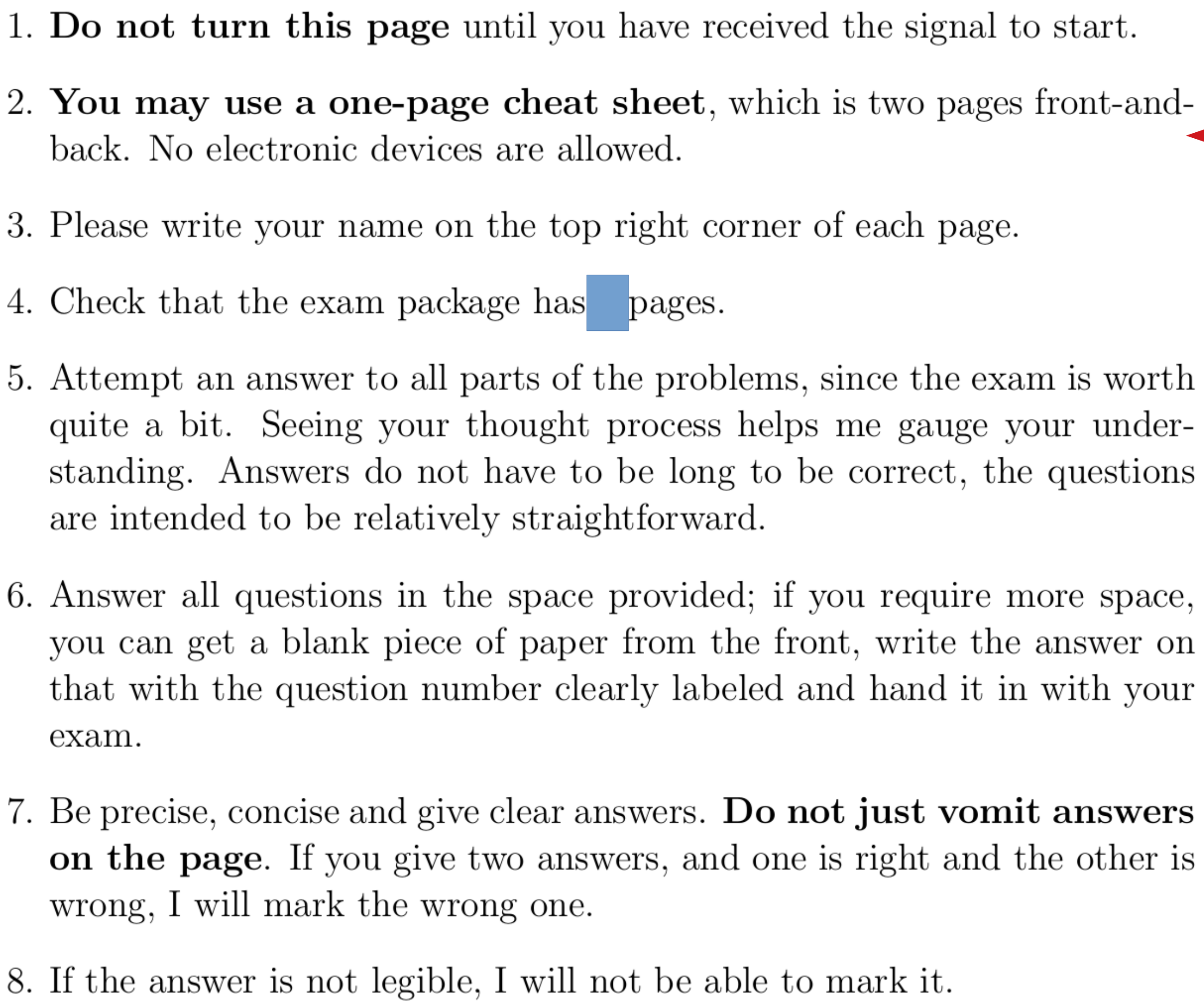
CMPUT 397: TD Control review

Rupam Mahmood

March 6, 2020

Midterm First page:

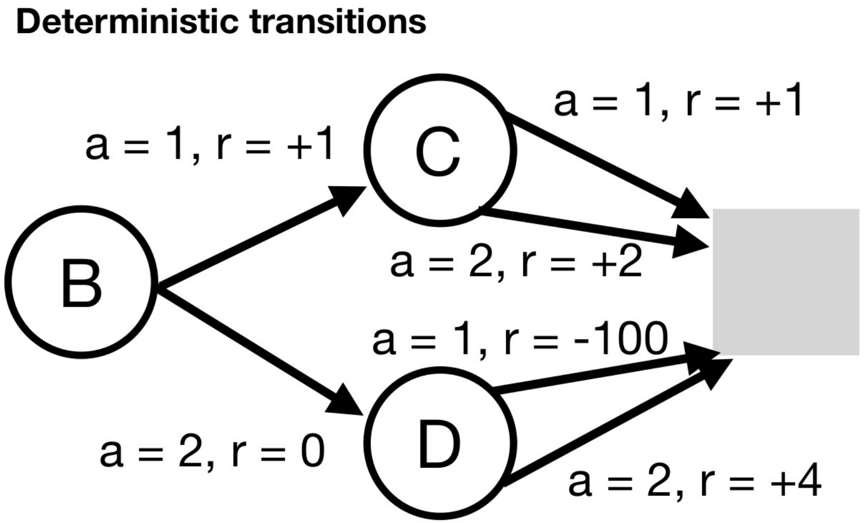


1. **Do not turn this page** until you have received the signal to start.
2. **You may use a one-page cheat sheet**, which is two pages front-and-back. No electronic devices are allowed.
3. Please write your name on the top right corner of each page.
4. Check that the exam package has [redacted] pages.
5. Attempt an answer to all parts of the problems, since the exam is worth quite a bit. Seeing your thought process helps me gauge your understanding. Answers do not have to be long to be correct, the questions are intended to be relatively straightforward.
6. Answer all questions in the space provided; if you require more space, you can get a blank piece of paper from the front, write the answer on that with the question number clearly labeled and hand it in with your exam.
7. Be precise, concise and give clear answers. **Do not just vomit answers on the page**. If you give two answers, and one is right and the other is wrong, I will mark the wrong one.
8. If the answer is not legible, I will not be able to mark it.

Hand-written only

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 $\epsilon = 0.1$.

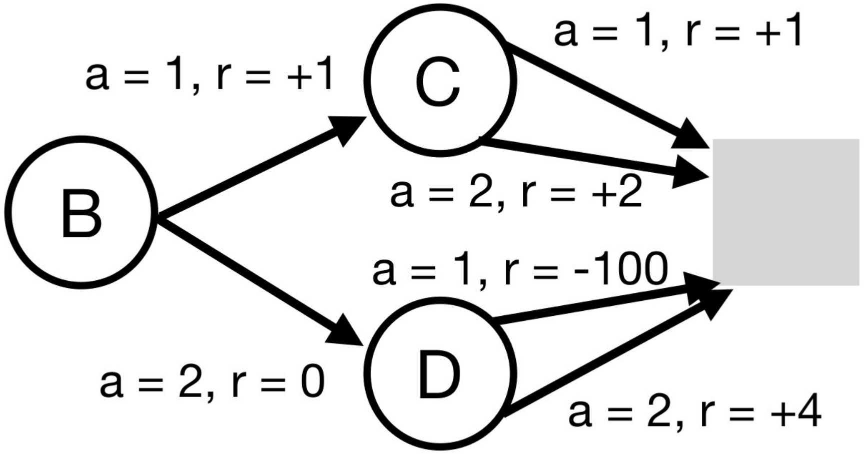
1. What is the optimal policy for this MDP and what are the action-values corresponding to the optimal policy: $q^*(s, a)$?



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 $\epsilon = 0.1$.

2. 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, assuming $\alpha = 0.1$? Start with state B, and perform the Sarsa update, then update the value of state D.

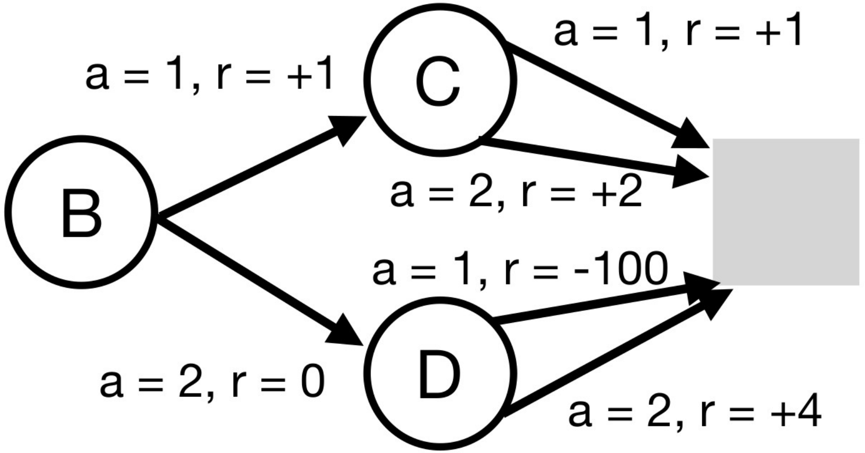
Deterministic transitions



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 $\epsilon = 0.1$.

2. 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, assuming $\alpha = 0.1$? Start with state B, and perform the Sarsa update, then update the value of state D.

3. Using the sample episode above, compute the updates Q-learning would make, with $\alpha = 0.1$. Again start with state B, and then state D.



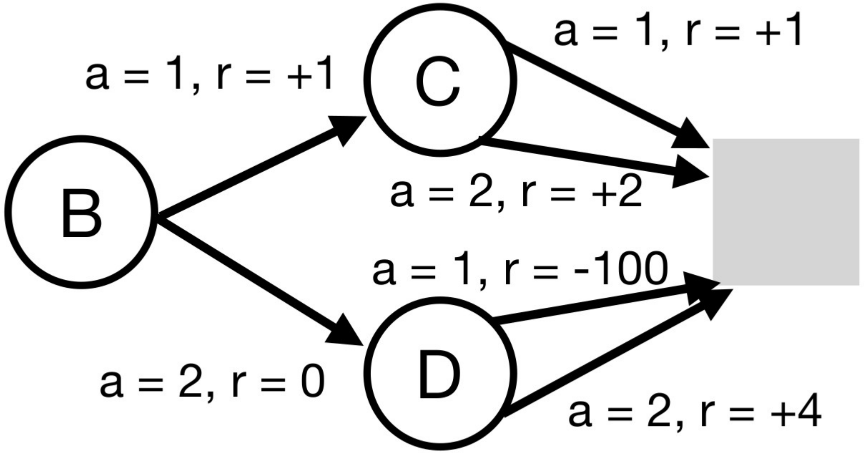
Deterministic transitions

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 $\epsilon = 0.1$.

2. 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, assuming $\alpha = 0.1$? Start with state B, and perform the Sarsa update, then update the value of state D.

4. 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?

Deterministic transitions

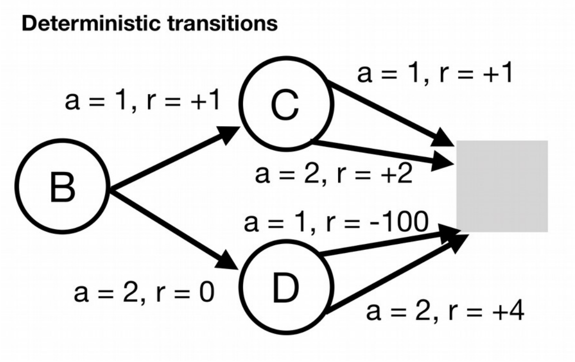


Consider the following MDP, with three states B, C and D ($\mathcal{S} = \{B, C, D\}$), and 2 actions ($\mathcal{A} = \{1, 2\}$), with $\gamma = 1.0$. Assume the action values are initialized $Q(s, a) = 0 \ \forall \ s \in \mathcal{S}$ and $a \in \mathcal{A}$. The agent takes actions according to an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

2. 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, assuming $\alpha = 0.1$? Start with state B, and perform the Sarsa update, then update the value of state D.

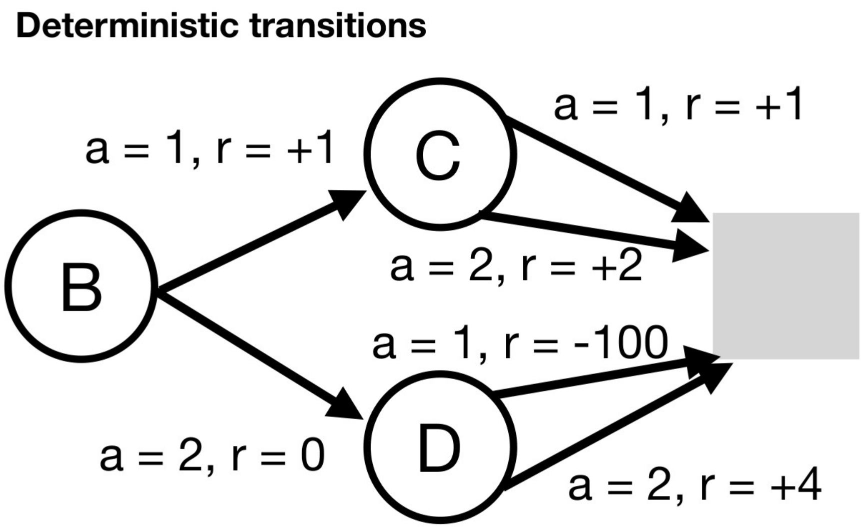
4. Let's consider one more episode: $S_0 = B, A_0 = 2, R_1 = 0, S_1 = D, A_1 = 1, R_2 = -100$. What would the Sarsa updates be? And what would the Q-learning updates be?

5. Assume you see one more episode, and it's the same on as in 4. Once more update the action values, for Sarsa and Q-learning. What do you notice?



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 $\epsilon = 0.1$.

6. What policy does Q-learning converge to? What policy does Sarsa converge to?



(*Exercise 6.12 S&B*) Suppose action selection is greedy. Is Q-learning then exactly the same algorithm as Sarsa? Will they make exactly the same action selections and weight updates? (**Additional Challenge:** What about Expected Sarsa? Does it have the same or different updates as Q-learning or Sarsa?)