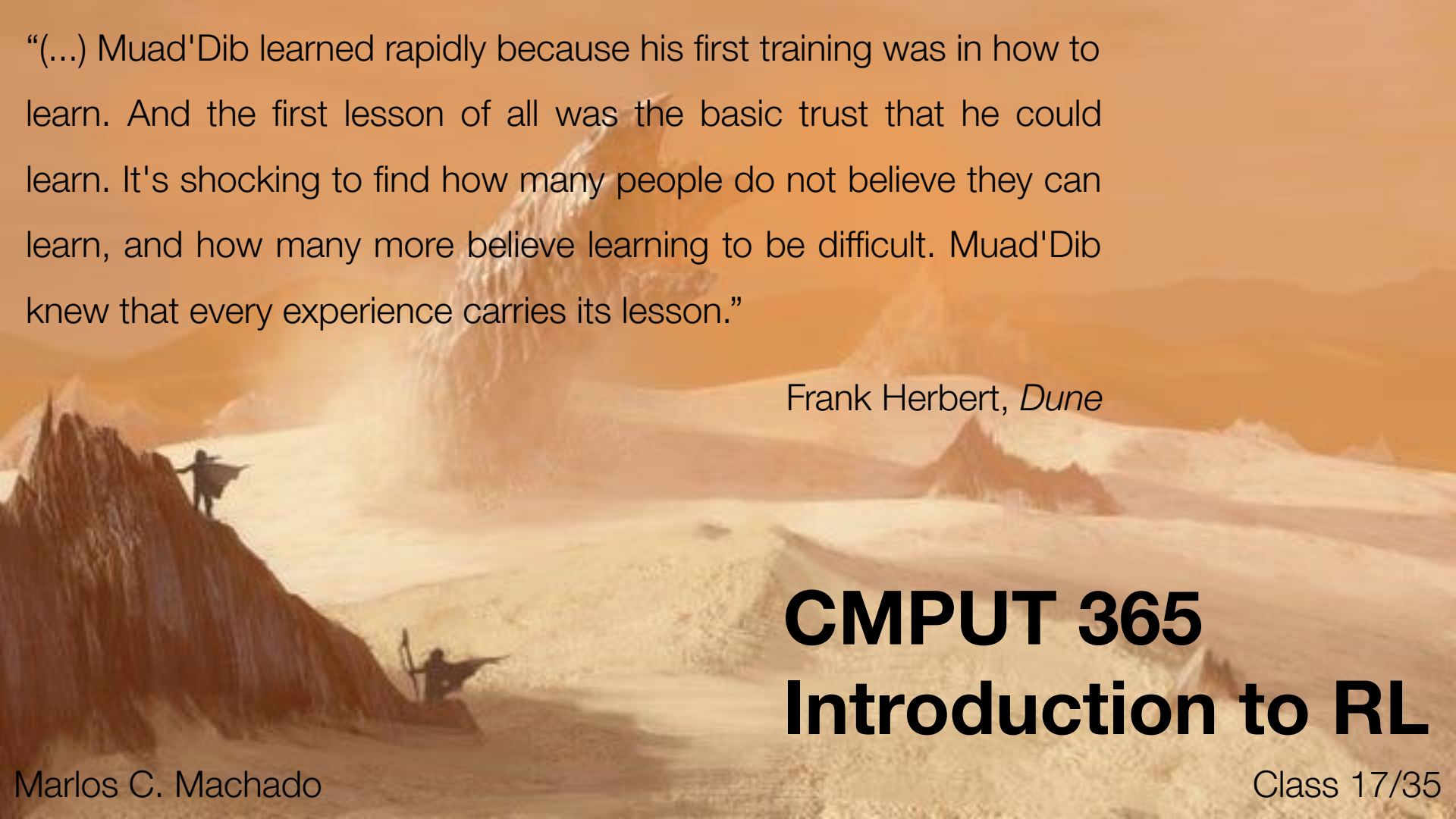

“(...) Muad'Dib learned rapidly because his first training was in how to learn. And the first lesson of all was the basic trust that he could learn. It's shocking to find how many people do not believe they can learn, and how many more believe learning to be difficult. Muad'Dib knew that every experience carries its lesson.”
Frank Herbert, *Dune*
CMPUT 365
Introduction to RL
Marlos C. Machado
Class 17/35

Coursera Reminder

You **should be enrolled in the private session** we created in Coursera for CMPUT 365.

I **cannot** use marks from the public repository for your course marks. You **need to check**, **every time**, if you are in the private session and if you are submitting quizzes and assignments to the private section.

At the end of the term, I **will not port grades** from the public session in Coursera.

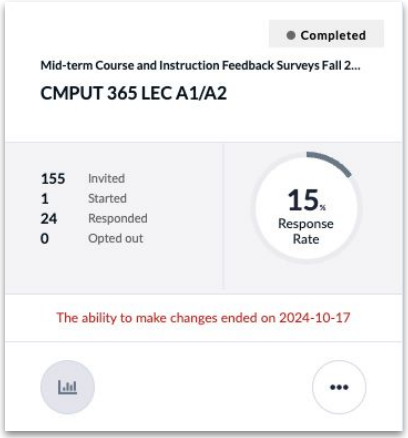
If you have any questions or concerns, **talk with the TAs** or email us `cmput365@ualberta.ca`.

Reminders and Notes

- Exam viewing:
  - It is happening today as well (
- What I plan to do today:
  - Finish overview of TD Learnin
- Useful information for you:
  - The Blackjack Programming A
  - The Programming Assignment for Temporal Difference Learning is due today.

Re-evaluation of midterm exams: Students will have access to their midterm exam during an exam viewing period. A student who has concerns about how specific questions of their midterm exam were marked can submit a request to the instructor via email within two weeks of the date they received their marked exam. The request should specify (1) which question is to be re-evaluated, (2) the rationale for such a request, and (3) the proposed marks. Importantly, once a request for re-evaluation is submitted, it is up to the instructor's discretion to adjust the marks. *Students won't be allowed to take their midterm exams with them, nor to take pictures of them*, so in case of concerns, the student is advised to take notes during the exam viewing period. The TAs are not authorized to weigh in on the midterm exams, this is something only the instructor can do. *Notice marks can also go down once a question is re-evaluated.*

SPOT: Mid-term Course Evaluation



Please, interrupt me at any time!

https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Chapter 6

Temporal-Difference Learning

Prediction

Why can we use $R_{t+1} + \gamma V(S_{t+1})$ instead of G_{t+1}?

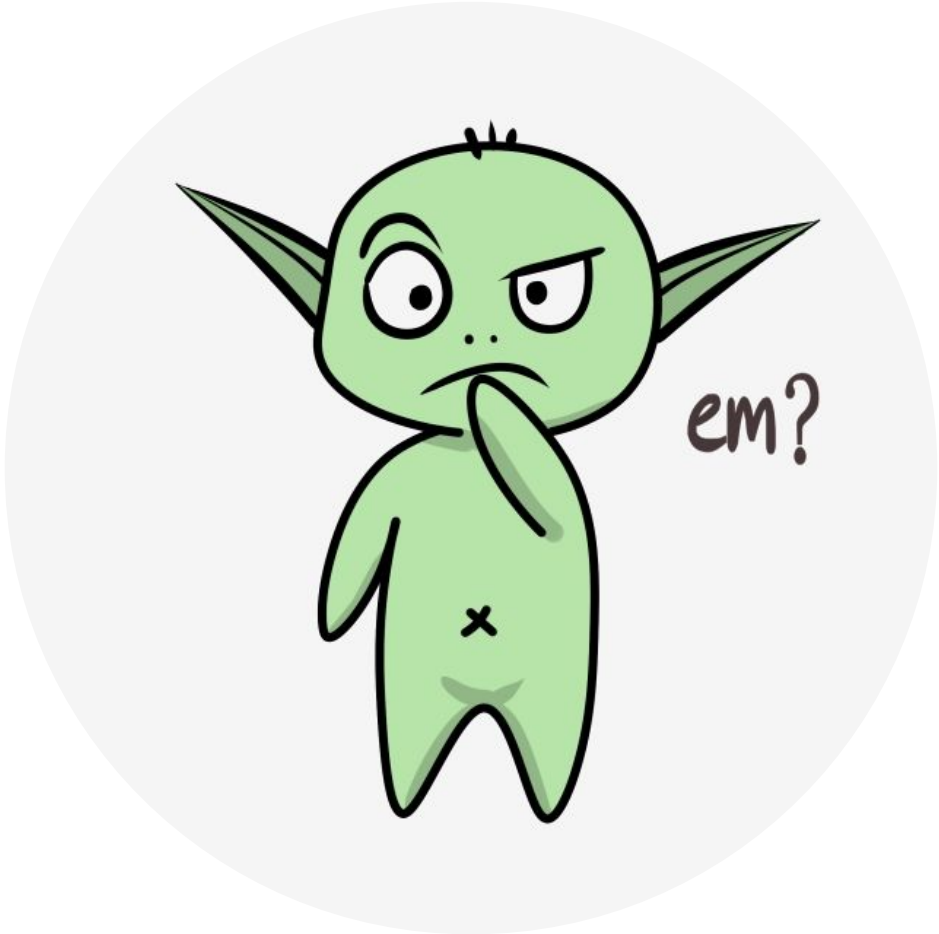
$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$$

$$G_t = R_{t+1} + \gamma G_{t+1}$$

$$v_\pi(s) = \mathbb{E}_\pi [G_t \mid S_t = s]$$

$$v_\pi(s) = \mathbb{E}_\pi [R_{t+1} + \gamma G_{t+1} \mid S_t = s]$$

$$v_\pi(s) = \mathbb{E}_\pi [R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s]$$



https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Optimality of TD(0)

- Under batch training, constant-α MC converges to values, V(s), that are sample averages of the actual returns experienced after visiting each state s. These are optimal estimates in the sense that they minimize the mean square error from the actual returns in the training set.

- But TD(0) gives us the answer that it is based on first modeling the Markov process and then computing the correct estimates given the model (the *certainty-equivalence estimate*).

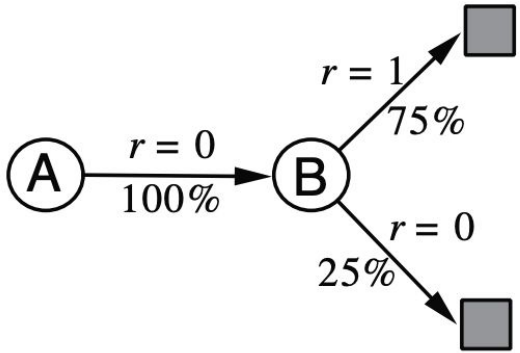
Example

Example 6.4: You are the Predictor Place yourself now in the role of the predictor of returns for an unknown Markov reward process. Suppose you observe the following eight episodes:

A, 0, B, 0	B, 1
B, 1	B, 1
B, 1	B, 1
B, 1	B, 0

V(A) = ?

V(B) = ?



Example

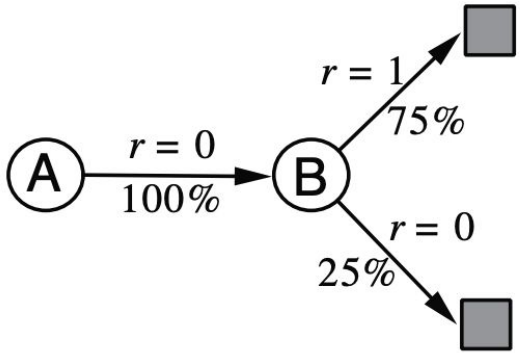
Example 6.4: You are the Predictor Place yourself now in the role of the predictor of returns for an unknown Markov reward process. Suppose you observe the following eight episodes:

A, 0, B, 0	B, 1
B, 1	B, 1
B, 1	B, 1
B, 1	B, 0

TD MC

V(A) = ? ¾ or 0?

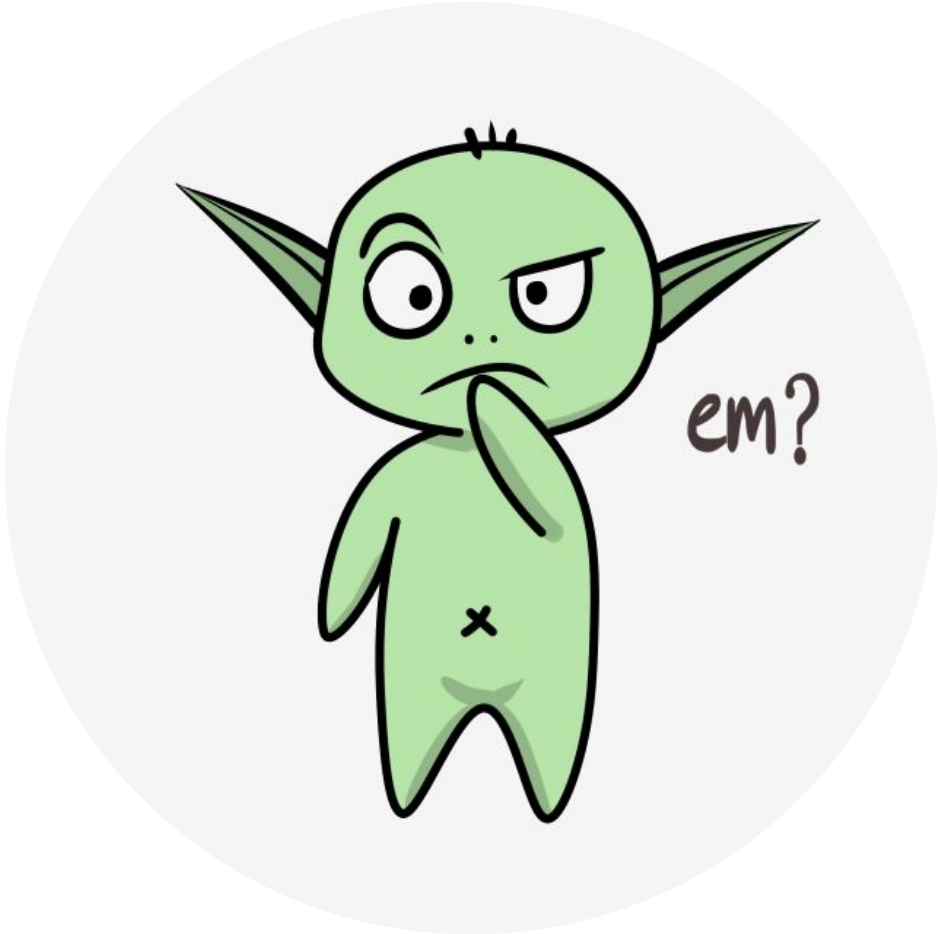
V(B) = ¾



TD vs Monte Carlo

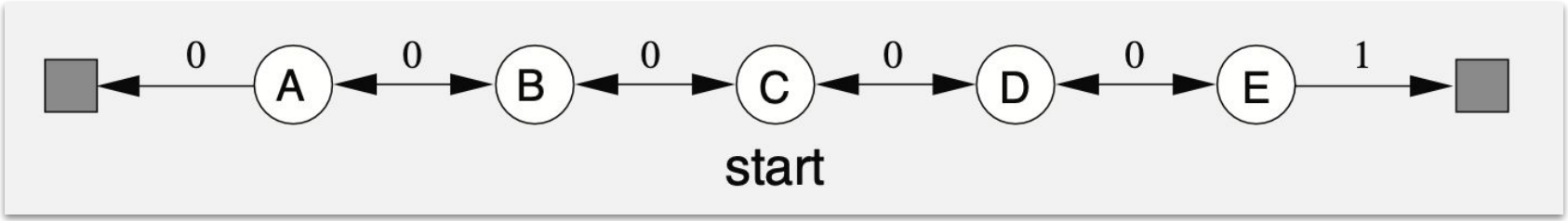
"Batch Monte Carlo methods always find the estimates that minimize mean square error on the training set, whereas batch TD(0) always finds the estimates that would be exactly correct for the maximum-likelihood model of the Markov process."

In general, the *maximum-likelihood estimate* of a parameter is the parameter value whose probability of generating the data is greatest.



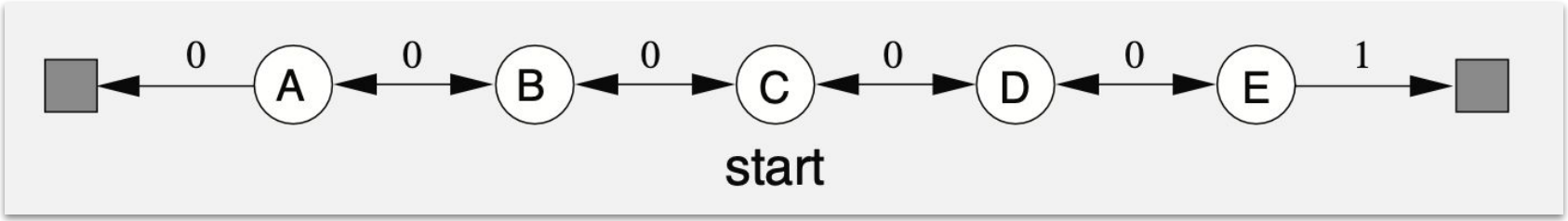
https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Example / Exercise: Random Walk



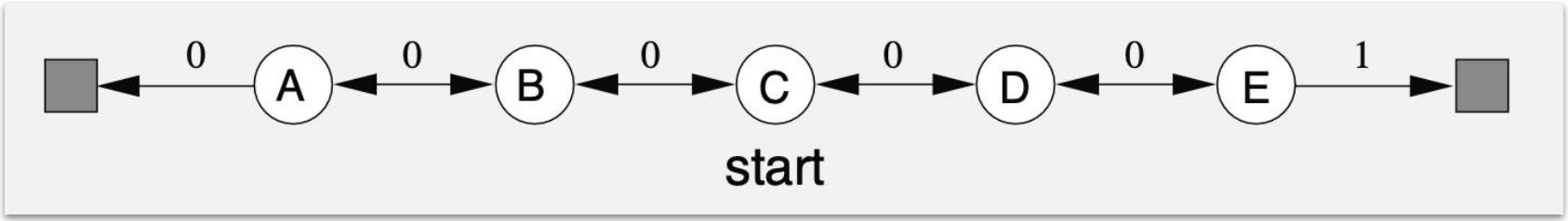
- *Markov reward process* (MRP), an MDP without actions.
- Start state: C
- One can go left or right, on each step, with equal probability.
- Reward: +1 on right exit, 0 otherwise.
- $\gamma = 1.0$
- All values are initialized with 0.5, that is $V(s) = 0.5$ for all s.

Example / Exercise: Random Walk

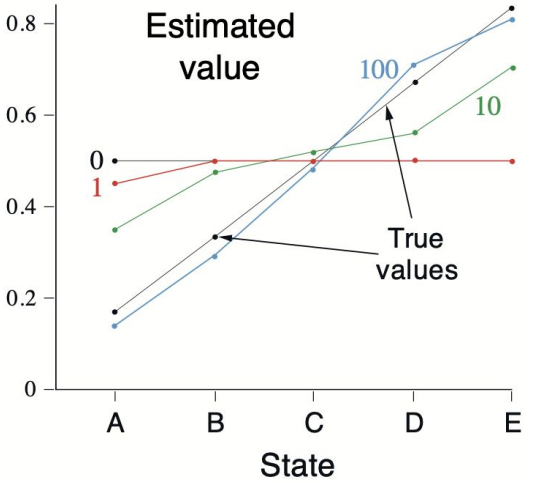


Q1. What does v_π encode in this problem?

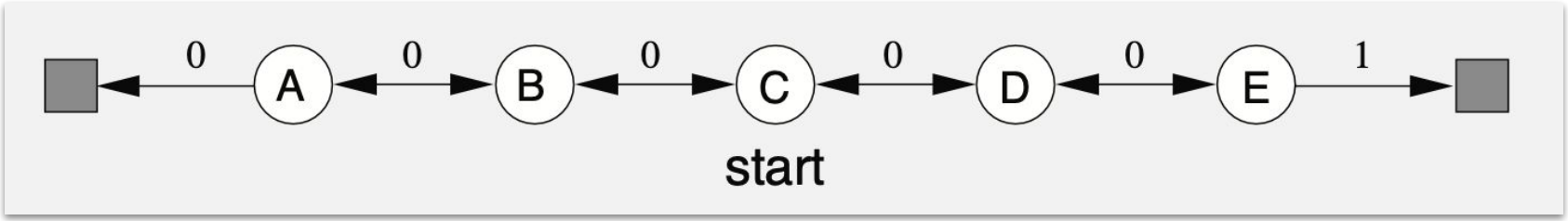
Example / Exercise: Random Walk



Q1. What does v_π encode in this problem?



Example / Exercise: Random Walk

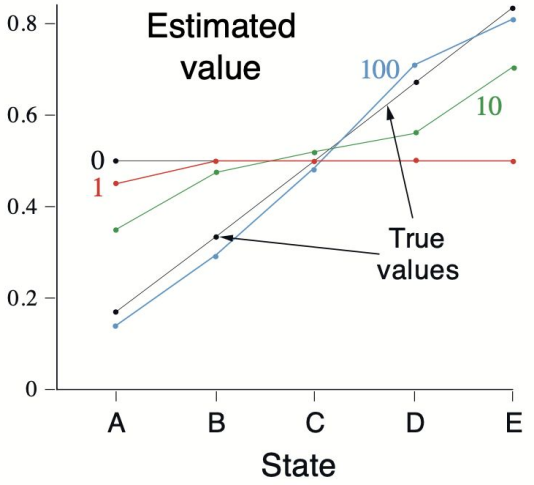


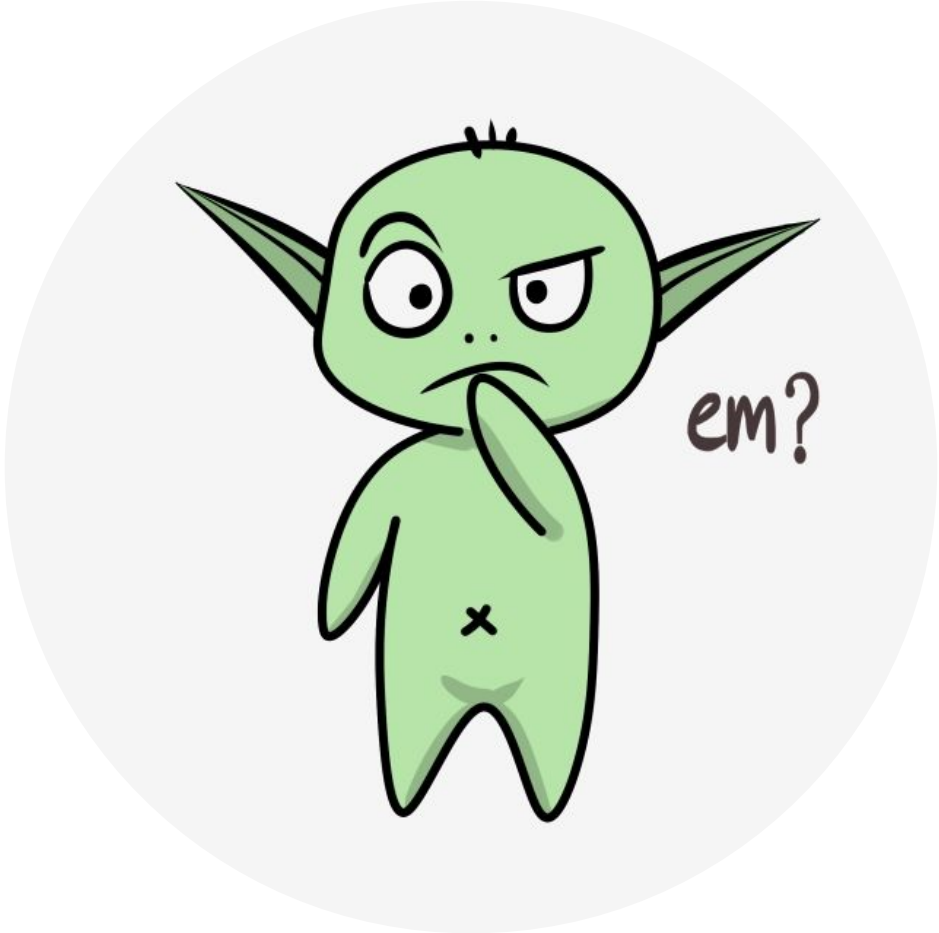
Q1. What does v_π encode in this problem?

Q2. The first episode results in a change in only V(A). What does this tell you about what happened on the first episode?

Q3. Why was only the estimate for this one state changed?

Q4. By exactly how much was it changed? (Assume α=0.1)





https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht