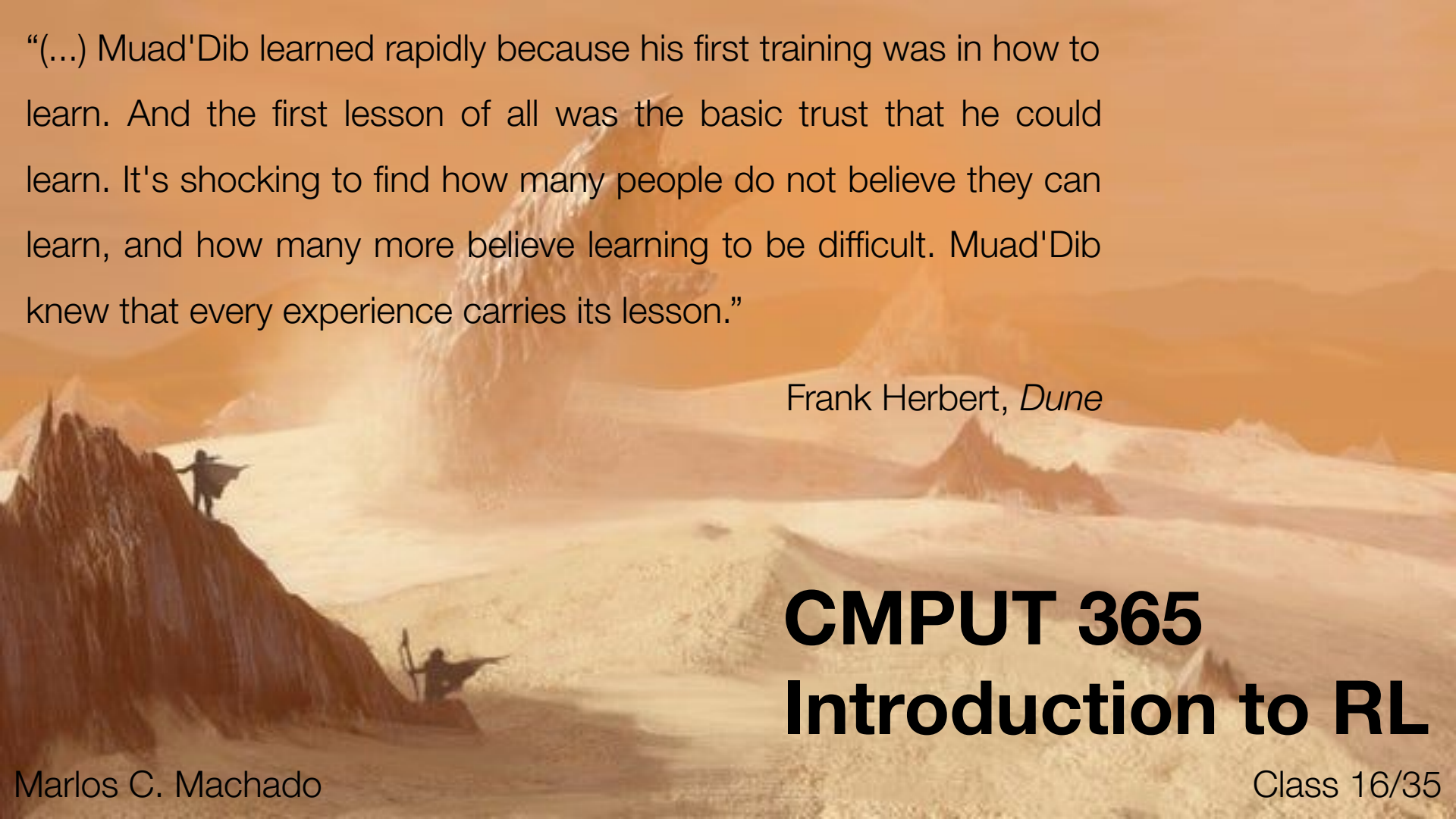

“(...) Muad'Dib learned rapidly because his first training was in how to learn. And the first lesson of all was the basic trust that he could learn. It's shocking to find how many people do not believe they can learn, and how many more believe learning to be difficult. Muad'Dib knew that every experience carries its lesson.”
Frank Herbert, *Dune*
CMPUT 365
Introduction to RL
Marlos C. Machado
Class 16/35

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`.

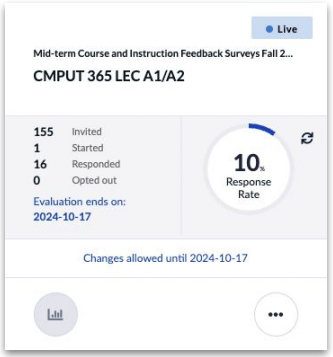
Reminders and Notes

- Exam viewing:
  - It will happen next Thursday (1 pm – 4pm) and Friday (2 pm – 5pm) at CSC 3-50.
- What I plan to do today:
  - Start overview of TD Learning M
- Useful information for you:
  - The Quiz for Temporal Difference
  - The Blackjack Programming Ass
  - The Programming Assignment fo

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 is Due Tomorrow

https://go.blueja.io/MlqAHuUezE-my_PTHx9IEg



Please, interrupt me at any time!

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

Chapter 6

Temporal-Difference Learning

Prediction

Temporal-difference learning – Why?

“If one had to identify one idea as central and novel to reinforcement learning, it would undoubtedly be temporal-difference (TD) learning.”

TD Prediction

A simple every-visit Monte Carlo method is:

$$V(S_t) \leftarrow V(S_t) + \alpha\Big[G_t - V(S_t)\Big]$$

What if we don't want to wait until we have a full return (end of episode)!

$$NewEstimate \leftarrow OldEstimate \; + \; StepSize \; \Big[Target - OldEstimate\Big]$$

TD Prediction

A simple every-visit Monte Carlo method is:

$$V(S_t) \leftarrow V(S_t) + \alpha\Big[\underbrace{G_t}_{\textbf{Target}} - V(S_t)\Big]$$

Temporal-Difference Learning:

$$V(S_t) \leftarrow V(S_t) + \alpha\Big[\underbrace{R_{t+1} + \gamma V(S_{t+1})}_{\textbf{Target}} - V(S_t)\Big]$$

TD Prediction

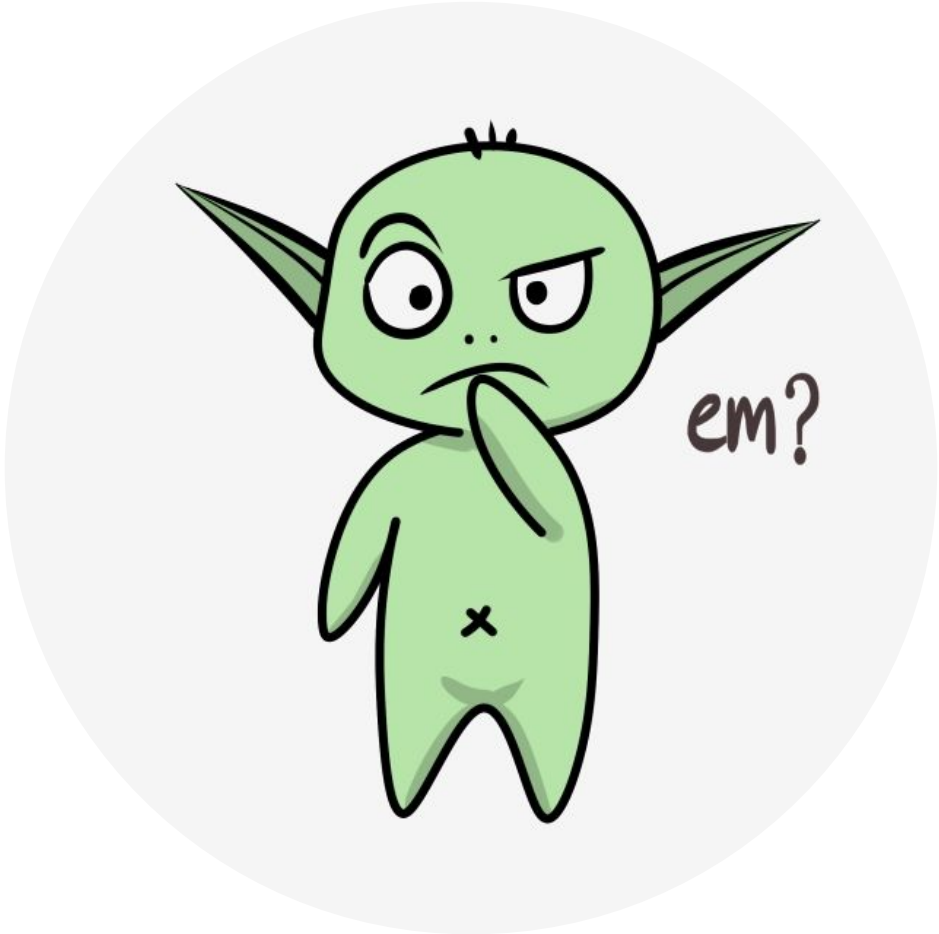
A simple every-visit Monte Carlo method is:

$$V(S_t) \leftarrow V(S_t) + \alpha\Big[G_t - V(S_t)\Big]$$

Temporal-Difference Learning (specifically, **one-step TD**, or **TD(0)**):

$$V(S_t) \leftarrow V(S_t) + \alpha\Big[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\Big]$$

These are estimates all the way down…



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

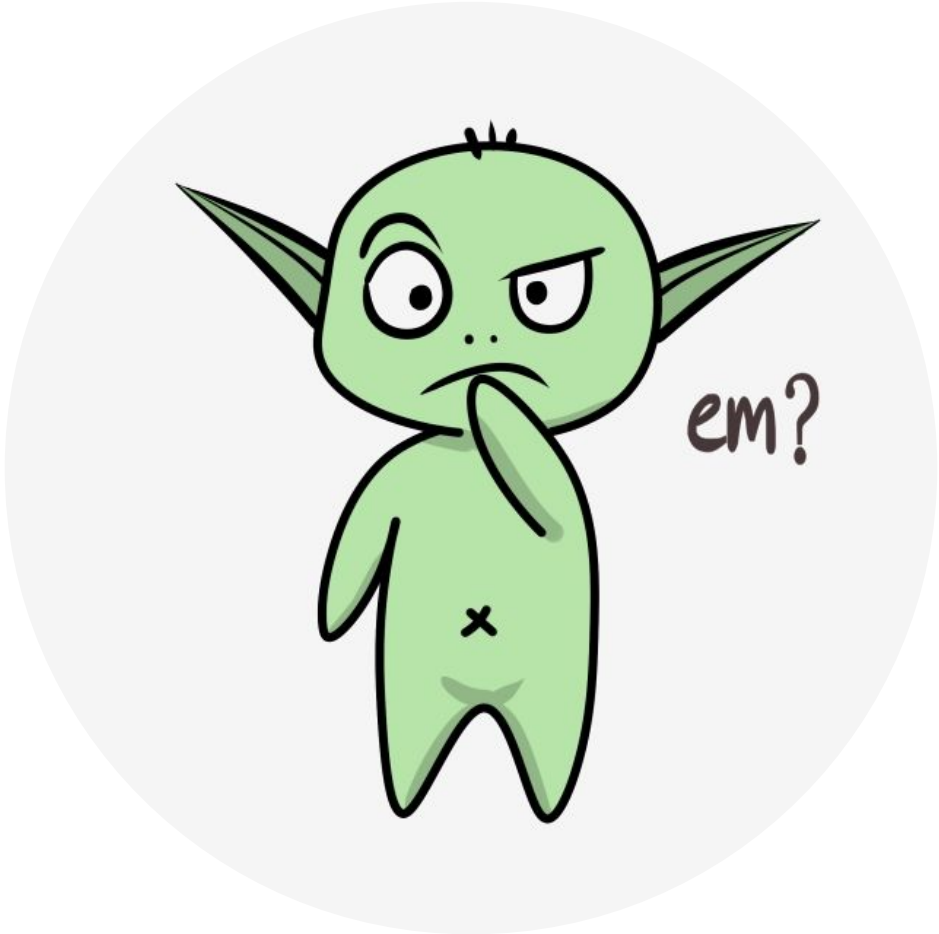
Tabular TD(0)

Tabular TD(0) for estimating v_π

Input: the policy π to be evaluated
Algorithm parameter: step size $\alpha \in (0, 1]$
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

Loop for each episode:
    Initialize S
    Loop for each step of episode:
        $A \leftarrow$ action given by π for S
        Take action A, observe R, S'
        $V(S) \leftarrow V(S) + \alpha\big[R + \gamma V(S') - V(S)\big]$
        $S \leftarrow S'$
    until S is terminal

Sample update



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

Temporal-Difference Error

$$\delta_t \doteq R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$$

Temporal-Difference Error

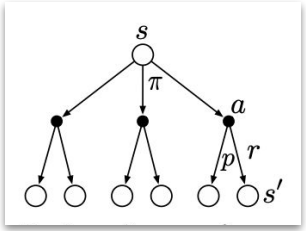
$$V(S_t) \leftarrow V(S_t) + \alpha\Big[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\Big]$$

$$\delta_t \doteq R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$$

TD is a sample update with bootstrapping

- Dynamic programming update:

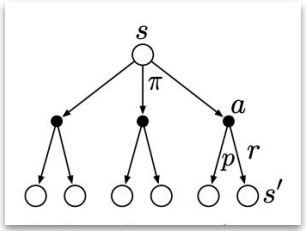
$$
\begin{aligned}
v_{k+1}(s) &\doteq \mathbb{E}_\pi[R_{t+1} + \gamma v_k(S_{t+1}) \mid S_t = s] \\
&= \sum_a \pi(a|s) \sum_{s',r} p(s', r \,|\, s, a) \Big[r + \gamma v_k(s') \Big]
\end{aligned}
$$



TD is a sample update with bootstrapping

- Dynamic programming update:

$$\begin{aligned} v_{k+1}(s) &\doteq \mathbb{E}_\pi[R_{t+1} + \gamma v_k(S_{t+1}) \mid S_t = s] \\ &= \sum_a \pi(a|s) \sum_{s',r} p(s', r | s, a) \Big[r + \gamma v_k(s') \Big] \end{aligned}$$



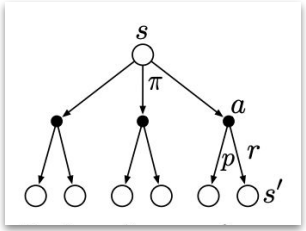
- Monte Carlo update:

$$V(S_t) \leftarrow V(S_t) + \alpha \Big[G_t - V(S_t) \Big]$$

TD is a sample update with bootstrapping

- Dynamic programming update:

$$\begin{aligned} v_{k+1}(s) &\doteq \mathbb{E}_\pi[R_{t+1} + \gamma v_k(S_{t+1}) \mid S_t = s] \\ &= \sum_a \pi(a|s) \sum_{s',r} p(s', r | s, a) \Big[r + \gamma v_k(s') \Big] \end{aligned}$$

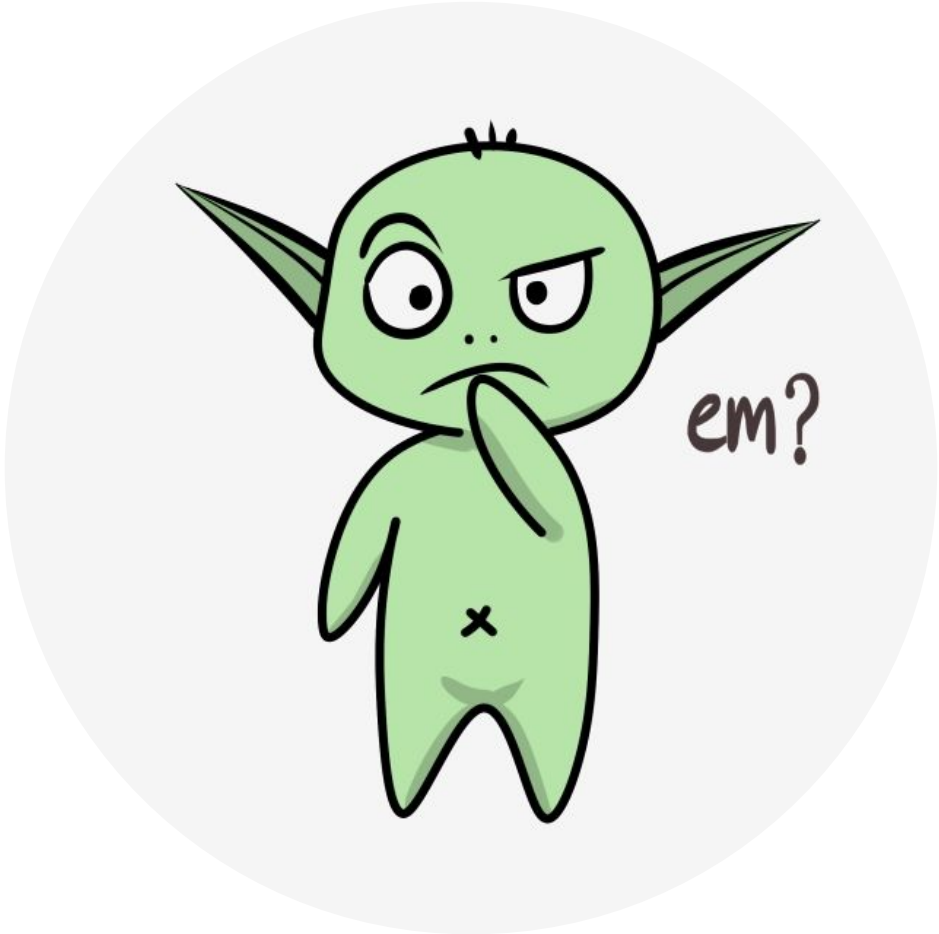


- Monte Carlo update:

$$V(S_t) \leftarrow V(S_t) + \alpha \Big[G_t - V(S_t) \Big]$$

- TD update:

$$V(S_t) \leftarrow V(S_t) + \alpha \Big[R_{t+1} + \gamma V(S_{t+1}) - V(S_t) \Big]$$



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

Example – Driving Home

Example 6.1: Driving Home Each day as you drive home from work, you try to predict how long it will take to get home. When you leave your office, you note the time, the day of week, the weather, and anything else that might be relevant. Say on this Friday you are leaving at exactly 6 o'clock, and you estimate that it will take 30 minutes to get home. As you reach your car it is 6:05, and you notice it is starting to rain. Traffic is often slower in the rain, so you reestimate that it will take 35 minutes from then, or a total of 40 minutes. Fifteen minutes later you have completed the highway portion of your journey in good time. As you exit onto a secondary road you cut your estimate of total travel time to 35 minutes. Unfortunately, at this point you get stuck behind a slow truck, and the road is too narrow to pass. You end up having to follow the truck until you turn onto the side street where you live at 6:40. Three minutes later you are home.

Example – Driving Home

The sequence of states, times, and predictions is thus as follows:

State	*Elapsed Time (minutes)*	*Predicted Time to Go*	*Predicted Total Time*
leaving office, friday at 6	0	30	30
reach car, raining	5	35	40
exiting highway	20	15	35
2ndary road, behind truck	30	10	40
entering home street	40	3	43
arrive home	43	0	43

Example – Driving Home

The rewards in this example are the elapsed times on each leg of the journey.[1] We are not discounting ($\gamma = 1$), and thus the return for each state is the actual time to go from that state. The value of each state is the *expected* time to go. The second column of numbers gives the current estimated value for each state encountered.

The sequence of states, times, and predictions is thus as follows:

State	*Elapsed Time (minutes)*	*Predicted Time to Go*	*Predicted Total Time*
leaving office, friday at 6	0	30	30
reach car, raining	5	35	40
exiting highway	20	15	35
2ndary road, behind truck	30	10	40
entering home street	40	3	43
arrive home	43	0	43

Example – Driving Home

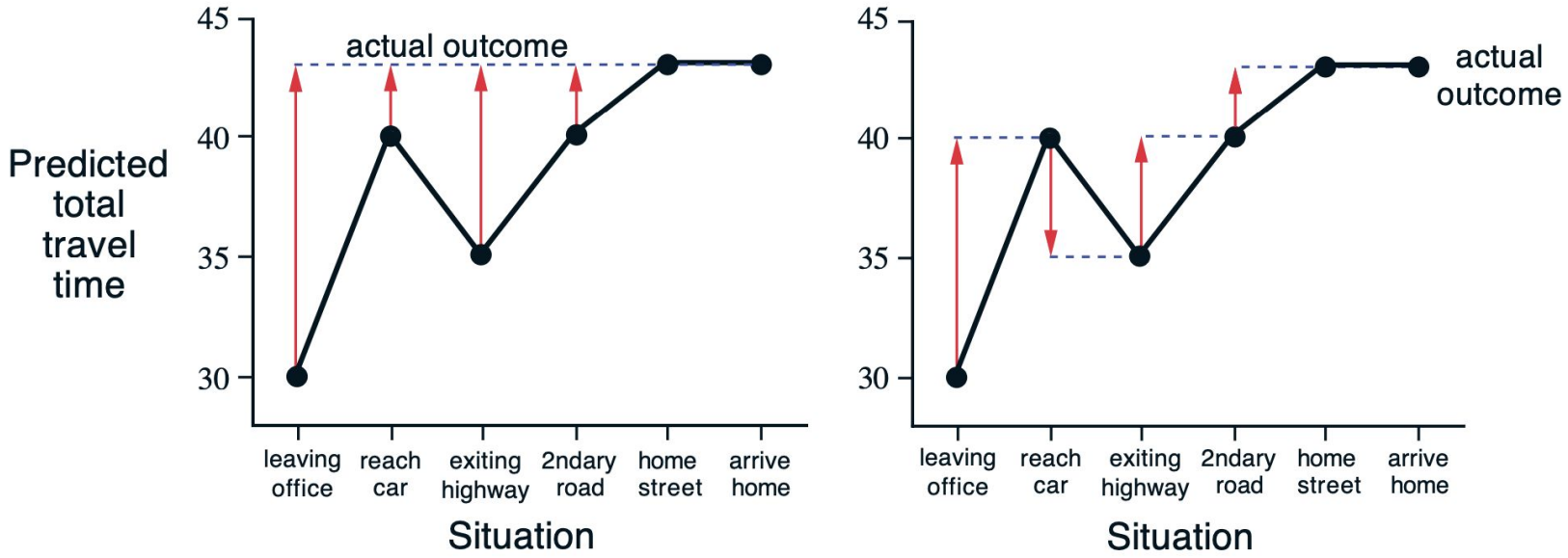
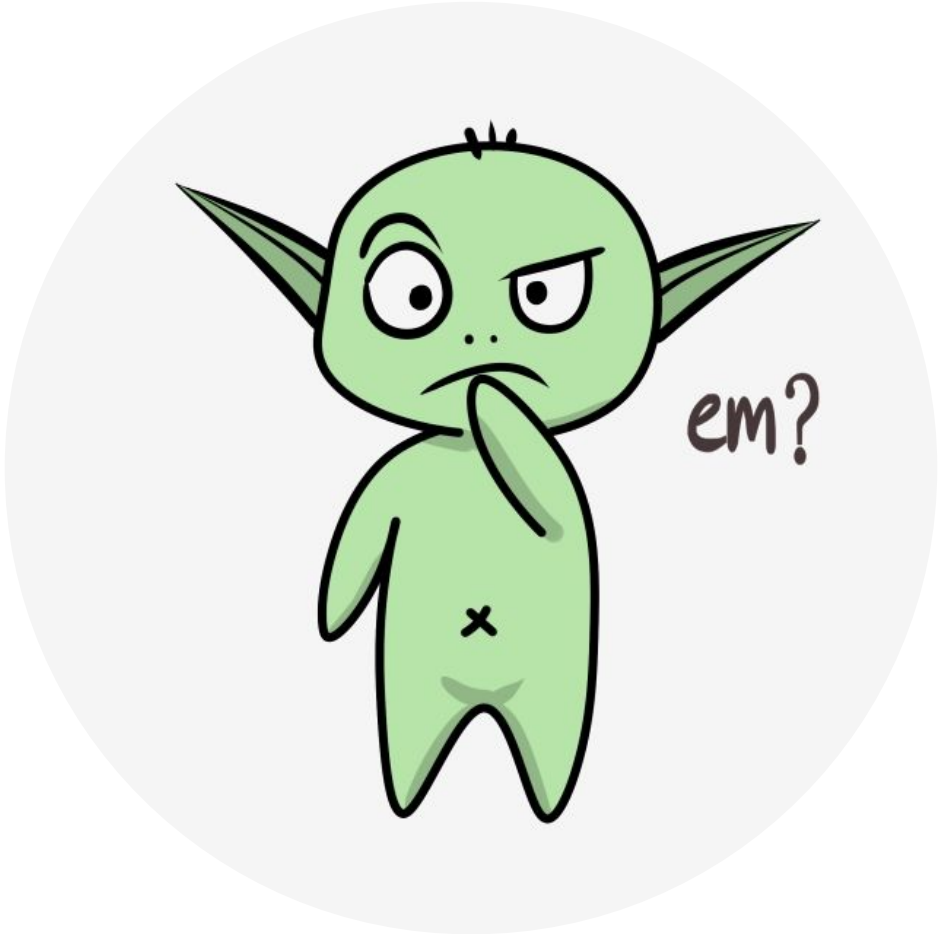


Figure 6.1: Changes recommended in the driving home example by Monte Carlo methods (left) and TD methods (right).



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