Reinforcement Learning: Tutorial 5

Temporal difference methods

Week 3
University of Amsterdam

Milena Kapralova
September 2024

Check-in

- How is it going?
- How is HW2?
- If you have any feedback so far, please mail me at *m.kapralova@uva.nl*

Outline

1. Admin
2. Temporal difference exercises
3. Ask anything about HW2

Admin

- Discrepancies between deadlines, from now on Wednesdays @ 17:00
- Any questions?

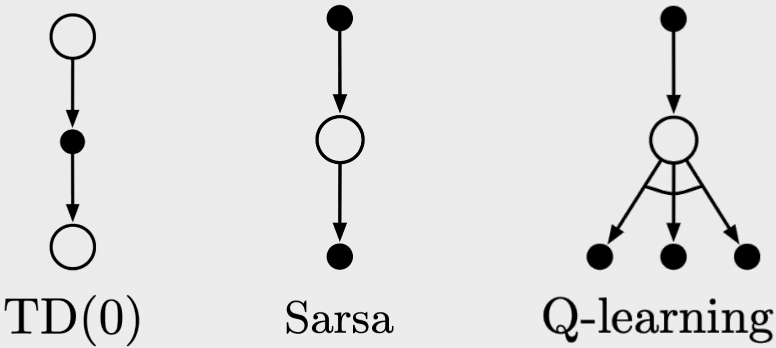
Tutorial 5 Overview

1. Temporal difference exercises
2. Ask anything about HW2

Tutorial 5 Overview

1. Temporal difference exercises
   - Questions 4.1-4.2
2. Ask anything about HW2

Theory Intermezzo: TD(0), SARSA, Q-learning



Q 4.1 Temporal difference learning (application)

Consider an undiscounted MDP with two states A and B, each with two possible actions 1 and 2, and a terminal state T with $V(T) = 0$. The transition and reward functions are unknown, but you have observed the following episode using a random policy:

- $A \xrightarrow[r_3=-3]{a_3=1} B \xrightarrow[r_4=4]{a_4=1} A \xrightarrow[r_5=-4]{a_5=2} A \xrightarrow[r_6=-3]{a_6=1} T$

1. What are the state(-action) value estimates $V(s)$ (or $Q(s,a)$) after observing the sample episode when applying
   a TD(0) (1-step TD)
   b SARSA
   c Q-learning

   where we initialize state(-action) values to 0 and use a learning rate $\alpha = 0.1$? Assume $\gamma = 1$.

Q 4.1 Temporal difference learning (application)

- $A \xrightarrow[r_3=-3]{a_3=1} B \xrightarrow[r_4=4]{a_4=1} A \xrightarrow[r_5=-4]{a_5=2} A \xrightarrow[r_6=-3]{a_6=1} T$
  Initialize state(-action) values to 0, $\alpha = 0.1$, $\gamma = 1$, $V(T) = 0$.

  a TD(0) (1-step TD)

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

Q 4.1 Temporal difference learning (application)

- $A \xrightarrow[r_3=-3]{a_3=1} B \xrightarrow[r_4=4]{a_4=1} A \xrightarrow[r_5=-4]{a_5=2} A \xrightarrow[r_6=-3]{a_6=1} T$
  Initialize state(-action) values to 0, $\alpha = 0.1$, $\gamma = 1$, $V(T) = 0$.

  a TD(0) (1-step TD)

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

$$\begin{aligned}
V(A) &= V(B) &&= 0 \\
V(A) &= 0 + 0.1 * (-3 + 0 - 0) &&= -0.3 \\
V(B) &= 0 + 0.1 * (4 + (-0.3) - 0) &&= 0.37 \\
V(A) &= -0.3 + 0.1 * (-4 + (-0.3) - (-0.3)) &&= -0.7 \\
V(A) &= -0.7 + 0.1 * (-3 + 0 - (-0.7)) &&= -0.930
\end{aligned}$$

Final:

$$V(A) = -0.930$$
$$V(B) = 0.37$$

Q 4.1 Temporal difference learning (application)

- $A \xrightarrow[r_3=-3]{a_3=1} B \xrightarrow[r_4=4]{a_4=1} A \xrightarrow[r_5=-4]{a_5=2} A \xrightarrow[r_6=-3]{a_6=1} T$
  Initialize state(-action) values to 0, $\alpha = 0.1$, $\gamma = 1$, $V(T) = 0$.

  b SARSA

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

Q 4.1 Temporal difference learning (application)

- $A \xrightarrow[r_3=-3]{a_3=1} B \xrightarrow[r_4=4]{a_4=1} A \xrightarrow[r_5=-4]{a_5=2} A \xrightarrow[r_6=-3]{a_6=1} T$
  Initialize state(-action) values to 0, $\alpha = 0.1$, $\gamma = 1$, $V(T) = 0$.

  b SARSA

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

$$\begin{aligned}
Q(A,1) = Q(A,2) = Q(B,1) = Q(B,2) &= 0 \\
Q(A,1) = 0 + 0.1 * (-3 + 0 - 0) &= -0.3 \\
Q(B,1) = 0 + 0.1 * (4 + 0 - 0) &= 0.4 \\
Q(A,2) = 0 + 0.1 * (-4 + (-0.3) - 0) &= -0.43 \\
Q(A,1) = -0.3 + 0.1 * (-3 + 0 - (-0.3)) &= -0.57
\end{aligned}$$

Final:

$$\begin{aligned}
Q(A,1) &= -0.57 & Q(A,2) &= -0.43 \\
Q(B,1) &= 0.4 & Q(B,2) &= 0
\end{aligned}$$

Q 4.1 Temporal difference learning (application)

- $A \xrightarrow[r_3=-3]{a_3=1} B \xrightarrow[r_4=4]{a_4=1} A \xrightarrow[r_5=-4]{a_5=2} A \xrightarrow[r_6=-3]{a_6=1} T$
  Initialize state(-action) values to 0, $\alpha = 0.1$, $\gamma = 1$, $V(T) = 0$.

  c Q-learning

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

Q 4.1 Temporal difference learning (application)

- $A \xrightarrow[r_3=-3]{a_3=1} B \xrightarrow[r_4=4]{a_4=1} A \xrightarrow[r_5=-4]{a_5=2} A \xrightarrow[r_6=-3]{a_6=1} T$
  Initialize state(-action) values to 0, $\alpha = 0.1$, $\gamma = 1$, $V(T) = 0$.

  c Q-learning

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

$$\begin{aligned}
Q(A,1) = Q(A,2) = Q(B,1) = Q(B,2) &= 0 \\
Q(A,1) = 0 + 0.1 * (-3 + 0 - 0) &= -0.3 \\
Q(B,1) = 0 + 0.1 * (4 + 0 - 0) &= 0.4 \\
Q(A,2) = 0 + 0.1 * (-4 + 0 - 0) &= -0.4 \\
Q(A,1) = -0.3 + 0.1 * (-3 + 0 - (-0.3)) &= -0.57
\end{aligned}$$

Final:

$$\begin{aligned}
Q(A,1) &= -0.57 & Q(A,2) &= -0.4 \\
Q(B,1) &= 0.4 & Q(B,2) &= 0
\end{aligned}$$

Q 4.2 Temporal difference learning (theory)

1. We can use Monte Carlo to get value estimates of a state with $V_M(S) = \frac{1}{M}\sum_{n=1}^{M} G_n(S)$ where $V_M(S)$ is the value estimate of state S after M visits of the state and $G_n(S)$ the return of an episode starting from S. Show that $V_M(S)$ can be written as the update rule $V_M(S) = V_{M-1}(S) + \alpha_M[G_M(S) - V_{M-1}(S)]$ and identify the learning rate α_M.

Q 4.2 Temporal difference learning (theory)

1. We can use Monte Carlo to get value estimates of a state with $V_M(S) = \frac{1}{M}\sum_{n=1}^{M} G_n(S)$ where $V_M(S)$ is the value estimate of state S after M visits of the state and $G_n(S)$ the return of an episode starting from S. Show that $V_M(S)$ can be written as the update rule $V_M(S) = V_{M-1}(S) + \alpha_M[G_M(S) - V_{M-1}(S)]$ and identify the learning rate α_M.

$$
\begin{aligned}
V_M(S) &= \frac{1}{M}\sum_{n=1}^{M} G_n(S) = \frac{1}{M}[G_M(S) + \frac{M-1}{M-1}\sum_{n=1}^{M-1} G_n(S)] \\
&= \frac{1}{M}[G_M(S) + (M-1)V_{M-1}(S)] \\
&= V_{M-1}(S) + \frac{1}{M}[G_M(S) - V_{M-1}(S)] \\
\rightarrow \alpha &= \frac{1}{M}
\end{aligned}
$$

Q 4.2 Temporal difference learning (theory)

2. Consider the TD-error

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

Q 4.2 Temporal difference learning (theory)

2. Consider the TD-error

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

   a. What is $\mathbb{E}[\delta_t | S_t = s]$ if δ_t uses the true state-value function V^π?

Q 4.2 Temporal difference learning (theory)

2. Consider the TD-error

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

a What is $\mathbb{E}[\delta_t | S_t = s]$ if δ_t uses the true state-value function V^π?

$$\begin{aligned}
\mathbb{E}[\delta_t | S_t = s] &= \mathbb{E}[R_{t+1} + \gamma V^\pi(S_{t+1}) - V^\pi(S_t) | S_t = s] && (1)\\
&= \mathbb{E}[R_{t+1} + \gamma V^\pi(S_{t+1}) | S_t = s] - V^\pi(s) && (2)\\
&= V^\pi(s) - V^\pi(s) && (3)\\
&= 0 && (4)
\end{aligned}$$

where the step from (2) to (3) follows from the Bellman equation.

Q 4.2 Temporal difference learning (theory)

2. Consider the TD-error

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

   b What is $\mathbb{E}[\delta_t | S_t = s, A_t = a]$ if δ_t uses the true state-value function V^π?

Q 4.2 Temporal difference learning (theory)

2. Consider the TD-error

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

   b. What is $\mathbb{E}[\delta_t | S_t = s, A_t = a]$ if δ_t uses the true state-value function V^π?

$$\begin{aligned}
\mathbb{E}[\delta_t | S_t = s, A_t = a] &= \mathbb{E}[R_{t+1} + \gamma V^\pi(S_{t+1}) - V^\pi(S_t) | S_t = s, A_t = a] \\
&= \mathbb{E}[R_{t+1} + \gamma V^\pi(S_{t+1}) | S_t = s, A_t = a] - V^\pi(s) \\
&= Q^\pi(s, a) - V^\pi(s) \\
&= A(s, a)
\end{aligned}$$

where $A(s, a)$ is the advantage function (important in later lectures).


WELL DONE!

Tutorial 5 Overview

1. Temporal difference exercises
2. Ask anything about HW2
   - Questions 3.4, 4.3

Ask anything about HW2

- 3.4: Coding (+ Little bit of theory)
- 4.3: Theory

That's it!

See you tomorrow