Sarsa

Shusen Wang

Derive TD Target

Discounted Return

Definition of discounted return:

- $U_t = R_t + \gamma \cdot R_{t+1} + \gamma^2 \cdot R_{t+2} + \gamma^3 \cdot R_{t+3} + \gamma^4 \cdot R_{t+4} + \cdots$

$$= \gamma \cdot (R_{t+1} + \gamma \cdot R_{t+2} + \gamma^2 \cdot R_{t+3} + \gamma^3 \cdot R_{t+4} + \cdots)$$

Discounted Return

- $U_t = R_t + \gamma \cdot R_{t+1} + \gamma^2 \cdot R_{t+2} + \gamma^3 \cdot R_{t+3} + \gamma^4 \cdot R_{t+4} + \cdots$

$$= R_t + \gamma \cdot \underbrace{(R_{t+1} + \gamma \cdot R_{t+2} + \gamma^2 \cdot R_{t+3} + \gamma^3 \cdot R_{t+4} + \cdots)}_{= U_{t+1}}$$

Discounted Return

Identity: $U_t = R_t + \gamma \cdot U_{t+1}.$

- $U_t = R_t + \gamma \cdot R_{t+1} + \gamma^2 \cdot R_{t+2} + \gamma^3 \cdot R_{t+3} + \gamma^4 \cdot R_{t+4} + \cdots$

$$= R_t + \gamma \cdot \underbrace{\left(R_{t+1} + \gamma \cdot R_{t+2} + \gamma^2 \cdot R_{t+3} + \gamma^3 \cdot R_{t+4} + \cdots\right)}_{= U_{t+1}}$$

Derive TD Target

Identity: $U_t = R_t + \gamma \cdot U_{t+1}$.

- Assume R_t depends on (S_t, A_t, S_{t+1}).
- $Q_\pi(s_t, a_t) = \mathbb{E}[U_t | s_t, a_t]$

Derive TD Target

Identity: $U_t = R_t + \gamma \cdot U_{t+1}$.

- Assume R_t depends on (S_t, A_t, S_{t+1}).
- $Q_\pi(s_t, a_t) = \mathbb{E}[U_t | s_t, a_t]$

$$= \mathbb{E}[R_t + \gamma \cdot U_{t+1} | s_t, a_t]$$

Derive TD Target

Identity: $U_t = R_t + \gamma \cdot U_{t+1}$.

- Assume R_t depends on (S_t, A_t, S_{t+1}).
- $Q_\pi(s_t, a_t) = \mathbb{E}[U_t | s_t, a_t]$

$$= \mathbb{E}[\underline{R_t + \gamma \cdot U_{t+1}} | s_t, a_t]$$

$$= \mathbb{E}[R_t | s_t, a_t] + \gamma \cdot \mathbb{E}[U_{t+1} | s_t, a_t]$$

Derive TD Target

Identity: $U_t = R_t + \gamma \cdot U_{t+1}$.

- Assume R_t depends on (S_t, A_t, S_{t+1}).
- $Q_\pi(s_t, a_t) = \mathbb{E}[U_t | s_t, a_t]$

$$= \mathbb{E}[R_t + \gamma \cdot U_{t+1} | s_t, a_t]$$

$$= \mathbb{E}[R_t | s_t, a_t] + \gamma \cdot \mathbb{E}[U_{t+1} | s_t, a_t]$$

$$\mathbb{E}[U_{t+1} | s_t, a_t] = \mathbb{E}[Q_\pi(S_{t+1}, A_{t+1}) | s_t, a_t]$$

Derive TD Target

Identity: $U_t = R_t + \gamma \cdot U_{t+1}$.

- Assume R_t depends on (S_t, A_t, S_{t+1}).
- $Q_\pi(s_t, a_t) = \mathbb{E}[U_t | s_t, a_t]$

$$= \mathbb{E}[R_t + \gamma \cdot U_{t+1} | s_t, a_t]$$

$$= \mathbb{E}[R_t | s_t, a_t] + \gamma \cdot \mathbb{E}[U_{t+1} | s_t, a_t]$$

$$\mathbb{E}[U_{t+1} | s_t, a_t] = \mathbb{E}[Q_\pi(S_{t+1}, A_{t+1}) | s_t, a_t]$$

Q_π eliminates all the future states and actions from time $t+2$.

Derive TD Target

- Assume R_t depends on (S_t, A_t, S_{t+1}).

- $Q_\pi(s_t, a_t) = \mathbb{E}[U_t | s_t, a_t]$

$$= \mathbb{E}[R_t + \gamma \cdot U_{t+1} | s_t, a_t]$$

$$= \mathbb{E}[R_t | s_t, a_t] + \gamma \cdot \mathbb{E}[U_{t+1} | s_t, a_t]$$

$$= \mathbb{E}[R_t | s_t, a_t] + \gamma \cdot \mathbb{E}[\underline{Q_\pi(S_{t+1}, A_{t+1})} | s_t, a_t].$$

Derive TD Target

Identity: $Q_\pi(s_t, a_t) = \mathbb{E}[R_t + \gamma \cdot Q_\pi(S_{t+1}, A_{t+1})]$, for all π.

Derive TD Target

Identity: $Q_\pi(s_t, a_t) = \mathbb{E}[R_t + \gamma \cdot Q_\pi(S_{t+1}, A_{t+1})]$, for all π.

- We do not know the expectation.
- Approximate it using Monte Carlo (MC).

Derive TD Target

Identity: $Q_\pi(s_t, a_t) = \mathbb{E}[R_t + \gamma \cdot Q_\pi(S_{t+1}, A_{t+1})]$, for all π.

y_t is its MC approximation.

- Let (s_{t+1}, r_t) be an observation of (S_{t+1}, R_t).
- Sample $a_{t+1} \sim \pi(\cdot \mid s_{t+1})$.
- TD target: $y_t = \underline{r_t} + \underline{\gamma \cdot Q_\pi(s_{t+1}, a_{t+1})}$.

Derive TD Target

Identity: $Q_\pi(s_t, a_t) = \mathbb{E}[R_t + \gamma \cdot Q_\pi(S_{t+1}, A_{t+1})]$, for all π.

y_t is its MC approximation.

TD learning: Encourage $Q_\pi(s_t, a_t)$ to approach y_t.

Sarsa: Tabular Version

Tabular Version

- We want to learn $Q_\pi(s, a)$.
- Suppose the numbers of states and actions are finite.
- Draw a table and learn the entries.

	Action a_1	Action a_2	Action a_3	Action a_4	$\cdots$
State s_1					
State s_2					
State s_3					
$\vdots$					

Sarsa (tabular version)

- Observe (s_t, a_t, r_t, s_{t+1}).
- Sample $a_{t+1} \sim \pi(\cdot \mid s_{t+1})$, where π is the policy function.
- TD target: $y_t = r_t + \gamma \cdot Q_\pi(s_{t+1}, a_{t+1})$.

	Action a_1	Action a_2	Action a_3	Action a_4	$\cdots$
State s_1					
State s_2					
State s_3					
$\vdots$					

Sarsa (tabular version)

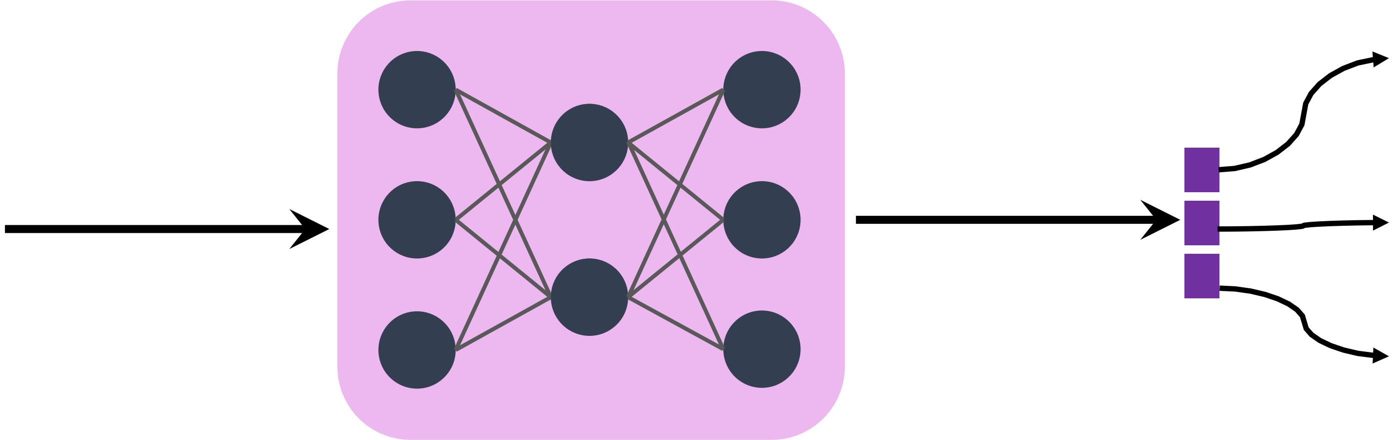
- Observe (s_t, a_t, r_t, s_{t+1}).
- Sample $a_{t+1} \sim \pi(\cdot \,|\, s_{t+1})$, where π is the policy function.
- TD target: $y_t = r_t + \gamma \cdot Q_\pi(s_{t+1}, a_{t+1})$.
- TD error: $\delta_t = Q_\pi(s_t, a_t) - y_t$.
- Update: $Q_\pi(s_t, a_t) \leftarrow \underbrace{Q_\pi(s_t, a_t) - \alpha \cdot \delta_t}_{\text{make } Q_\pi(s_t, a_t) \text{ closer to } y_t}$.

Sarsa: Neural Network Version

Value Network Version

- Approximate $Q_\pi(s, a)$ by the value network, $q(s, a|\mathbf{w})$.

state s



$q(s, \text{“left”}; \mathbf{w})$

$q(s, \text{“right”}; \mathbf{w})$

$q(s, \text{“up”}; \mathbf{w})$

Value Network

(parameterized by $\mathbf{w}$)

Value Network Version

- Approximate $Q_\pi(s, a)$ by the value network, $q(s, a|\mathbf{w})$.
- Note that $Q_\pi(s, a)$ and $q(s, a|\mathbf{w})$ depend on π.
- q is used as the critic who evaluates the actor. (Actor-Critic Method.)
- We want to learn the parameter, $\mathbf{w}$.

Sarsa (Value Network Version)

- Observe (s_t, a_t, r_t, s_{t+1}).
- Sample $a_{t+1} \sim \pi(\cdot\,|s_{t+1})$, where π is the policy function.
- TD target: $y_t = \underline{r_t} + \gamma \cdot \underline{q(s_{t+1}, a_{t+1}|\mathbf{w})}$.

Sarsa (Value Network Version)

- Observe (s_t, a_t, r_t, s_{t+1}).
- Sample $a_{t+1} \sim \pi(\cdot \mid s_{t+1})$, where π is the policy function.
- TD target: $y_t = r_t + \gamma \cdot q(s_{t+1}, a_{t+1} \mid \mathbf{w})$.
- TD error: $\delta_t = q(s_t, a_t \mid \mathbf{w}) - y_t$.
- SGD: $\mathbf{w} \leftarrow \mathbf{w} - \alpha \cdot \boxed{\delta_t \cdot \frac{\partial\, q(s_t, a_t \mid \mathbf{w})}{\partial\, \mathbf{w}}}$.

Summary

- **Goal:** Learn the action-value function Q_π.
- **Tabular version** (directly learn Q_π).
  - There are finite states and actions.
  - Draw a table, and update the table using Sarsa.
- **Value network version** (function approximation).
  - Approximate Q_π by the value network $q(s, a|\mathbf{w})$.
  - Update the parameter, $\mathbf{w}$, using Sarsa.
  - Application: actor-critic method.

Thank you!