Basic RL.2

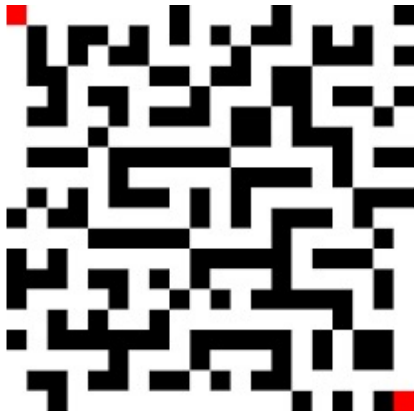
Judy Tutorial

“the goal of an RL agent is to find an ***optimal policy*** that maximizes the ***expected return***”

Episodes

When the agent-env interation breaks naturally into subsequences

- Plays of a game
- Trips trough a maze
- Any sort of repeated interaction



Start / Terminal state

Each episode starts at a **starting state**, ends in a **terminal state**,

starting state can be a sample from a standard **distribution of starting state**

History

Time horizon: T

T = a fixed num, Finite and Episodic

$$h^T = [s_0, a_0, r_0, s_1, \mathrm{a}_1, \mathrm{r}_1, \ldots \underline{\boldsymbol{s_t}, \boldsymbol{a_t}, \boldsymbol{r_t}, \boldsymbol{s_{t+1}}}, \ldots s_T, a_T, r_T, s_{T+1}]$$

$T = \infty$, Infinite and Continuous

$$h^\infty = [s_0, a_0, r_0, s_1, \ldots \boldsymbol{s_t}, \boldsymbol{a_t}, \boldsymbol{r_t}, \boldsymbol{s_{t+1}}, \ldots \ldots]$$

Return

$$R_t \triangleq r_t + r_{t+1} + r_{t+2} + \cdots + r_T = \sum_{i=0}^{T} r_{t+i}$$

$$R_t \triangleq r_t + r_{t+1} + r_{t+2} + \cdots = \sum_{i=0}^{T=\infty} r_{t+i}$$

Discounting
$\gamma \in [\mathbf{0}, \mathbf{1}]$

$$R_t \triangleq r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots = \sum_{i=0}^{\boldsymbol{T}=\infty} \gamma^i r_{t+i}$$

Discounting
$\boldsymbol{\gamma \in [0, 1]}$

- $\gamma = 0$

Myopic – only concerned with immediate rewards

$$R_t = r_t$$

- $\gamma = 0.5$

$$R_t = r_t + 0.5r_{t+1} + 0.25r_{t+2} + 0.125r_{t+3} + 0.0625r_{t+4} \ldots$$

- $\gamma = 0.9$

Farsighted – concerned with more about future rewards

$$R_t = r_t + 0.9r_{t+1} + 0.81r_{t+2} + 0.729r_{t+3} + 0.6561r_{t+4} \ldots$$

Value function

Expected Return

Future rewards that can be expected

How good it is for an agent to be in a state s or a state-action pair (s, a) following a specific policy $\pi(a|s)$

State value function $V(s)$ for policy π

for all $s \in \mathcal{S}$

Expected Return

$$V_\pi(s) \triangleq \underline{\mathbb{E}_\pi[R_t | s_t = s]}$$

$$= \mathbb{E}_\pi\left[\sum_{i=0}^{T} \gamma^i r_{t+i} | s_t = s\right]$$

Starting point

Action value function $Q(s, a)$ for policy π

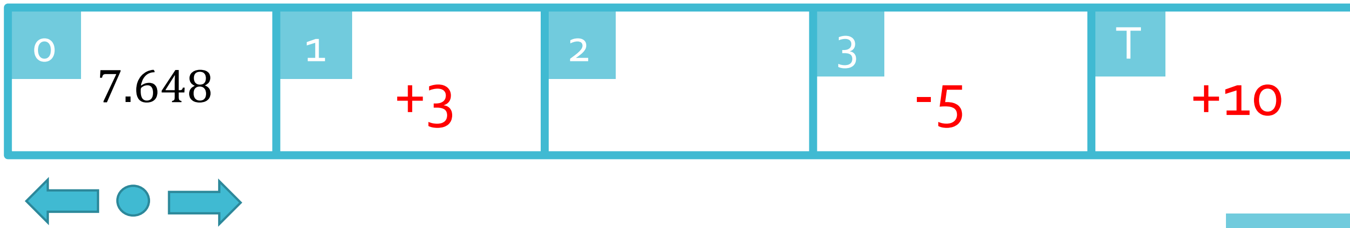
for all $s \in \mathcal{S}, a \in \mathcal{A}$

Expected Return

$$Q_\pi(s, a) \triangleq \mathbb{E}_\pi[R_t | s_t = s, a_t = a]$$
$$= \mathbb{E}_\pi\left[\sum_{i=0}^{T} \gamma^i r_{t+i} | s_t = s, a_t = a\right]$$

Starting point

Chain (env)



$\gamma = 0.9$

$\mathcal{S}$: (5,)

$\mathcal{A}$: (3,) – left(l), stay(s), right(r)

π_u: $\pi(l|s) = \frac{1}{3}, \pi(s|s) = \frac{1}{3}, \pi(r|s) = \frac{1}{3}$ <- uniform random policy

$p(s'|s,a)$: deterministic

Given 2 history trajs with $T = 10$, what is $V_{\pi_u}(s_0)$?

History Traj 0: (0,r) -> (1,r) -> (2,r) -> (3,r) -> T

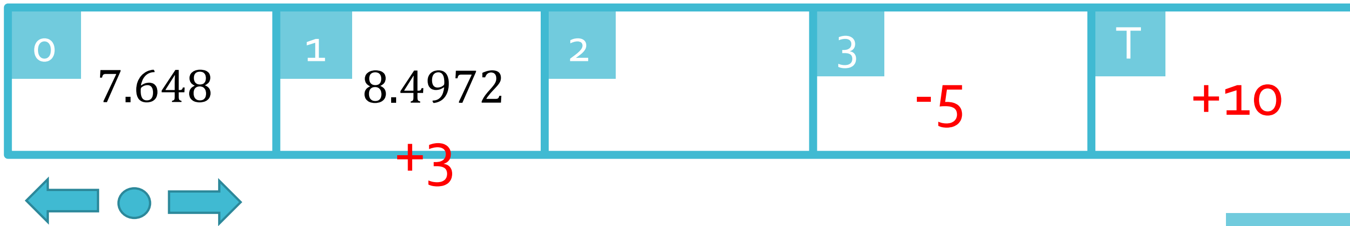
$$R_{h_0}(s_0) = 0 + \gamma(+3) + \gamma^2 0 + \gamma^3(-5) + \gamma^4(+10) = 5.616$$

History Traj 1 : (0,r) -> (1,s) -> (1,s) -> (1,s) -> (2,r) -> (3,r) -> T

$$R_{h_1}(s_0) = 0 + \gamma(+3) + \gamma^2(+3) + \gamma^3(+3) + \gamma^4 0 + \gamma^5(-5) + \gamma^6(+10) = 9.68$$

$$V_{\pi_u}(s_0) = \frac{1}{2}\left[R_{h_0}(s_0) + R_{h_1}(s_0)\right] = 7.648$$

*To illustrate what a state value can be, we simply average the returns over trajs for estimating the value. There are other ways to calculate this value in terms of the knowledge of dynamics and the way you collect trajs



$\gamma = 0.9$

$\mathcal{S}$: (5,)
$\mathcal{A}$: (3,) – left(l), stay(s), right(r)
π_u: $\pi(l|s) = \frac{1}{3}, \pi(s|s) = \frac{1}{3}, \pi(r|s) = \frac{1}{3}$ uniform random policy
$p(s'|s,a)$: deterministic

Given 2 history trajs with $T = 10$, what is $V_{\pi_u}(s_1)$?

History Traj 0: (0,r) -> (1,r) -> (2,r) -> (3,r) -> T

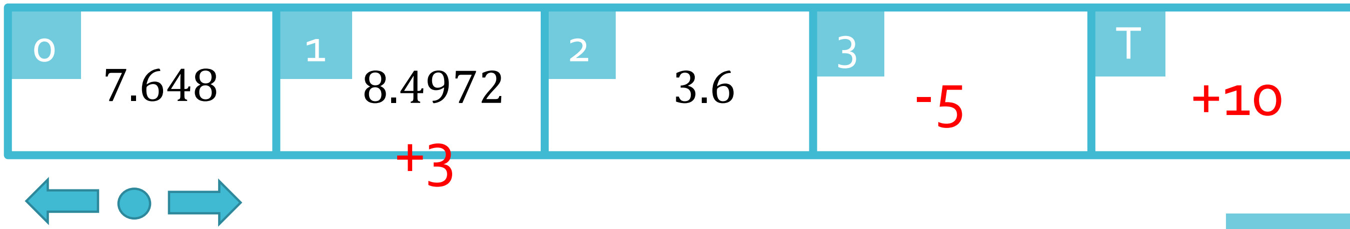
$$R_{h_0}(s_1) = 3 + \gamma 0 + \gamma^2(-5) + \gamma^3(+10) = 6.24$$

History Traj 1: (0,r) -> (1,s) -> (1,s) -> (1,s) -> (2,r) -> (3,r) -> T

$$R_{h_1}(s_1) = 3 + \gamma(+3) + \gamma^2(+3) + \gamma^3 0 + \gamma^4(-5) + \gamma^5(+10) = 10.7544$$

$$V_{\pi_u}(s_1) = \frac{1}{2}\left[R_{h_0}(s_1) + R_{h_1}(s_1)\right] = 8.4972$$

Chain (env)



$\gamma = 0.9$

$\mathcal{S}$: (5,)
$\mathcal{A}$: (3,) – left(l), stay(s), right(r)
π_u: $\pi(l|s) = \frac{1}{3}, \pi(s|s) = \frac{1}{3}, \pi(r|s) = \frac{1}{3}$ uniform random policy
$p(s'|s,a)$: deterministic

Given 2 history trajs with $T = 10$, what is $V_{\pi_u}(s_2)$?

History Traj 0: (0,r) -> (1,r) -> (2,r) -> (3,r) -> T

$$R_{h_0}(s_2) = 0 + \gamma(-5) + \gamma^2(+10) = 3.6$$

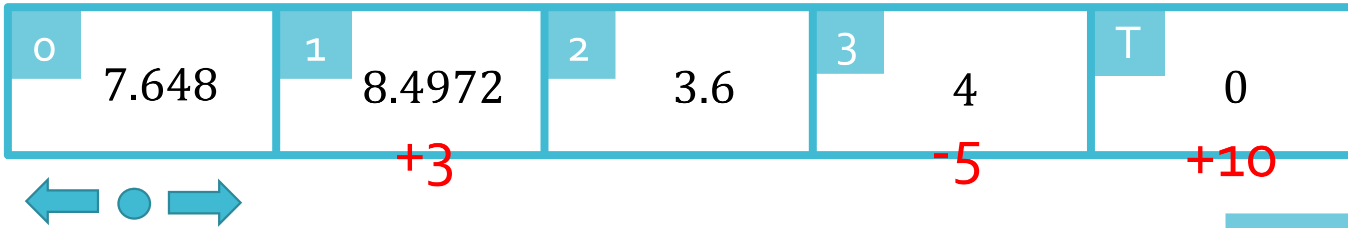
History Traj 1: (0,r) -> (1,s) -> (1,s) -> (1,s) -> (2,r) -> (3,r) -> T

$$R_{h_1}(s_2) = R_{h_0}(s_2)$$

$$V_{\pi_u}(s_2) = \frac{1}{2}\left[R_{h_0}(s_2) + R_{h_1}(s_2)\right] = 3.6$$

Chain (env)

Note: Value for the terminal state = 0



$\gamma = 0.9$

$\mathcal{S}$: (5,)

$\mathcal{A}$: (3,) – left(l), stay(s), right(r)

π_u: $\pi(l|s) = \frac{1}{3}$, $\pi(s|s) = \frac{1}{3}$, $\pi(r|s) = \frac{1}{3}$ uniform random policy

$p(s'|s,a)$: deterministic

Given 2 history trajs with $T = 10$, what is $V_{\pi_u}(s_3)$?

History Traj 0: (0,r) -> (1,r) -> (2,r) -> (3,r) -> T

$$R_{h_0}(s_3) = -5 + \gamma(+10) = 4$$

History Traj 1: (0,r) -> (1,s) -> (1,s) -> (1,s) -> (2,r) -> (3,r) -> T

$$R_{h_1}(s_3) = R_{h_0}(s_3)$$

$$V_{\pi_u}(s_3) = \frac{1}{2}\left[R_{h_0}(s_3) + R_{h_1}(s_3)\right] = 4$$

Optimal policy π^*

- $\pi \geq \pi'$ iff $V_\pi(s) \geq V_{\pi'}(s)$, for all $s \in \mathcal{S}$

- There is always at least one policy that is better than or euqal to all other policies

- $\pi^* = \mathrm{argmax}_\pi V_\pi(s)$

Or

- $\pi^* = \mathrm{argmax}_\pi Q_\pi(s, a)$

Optimal value V^*, Q^*

- $V^*(s) \triangleq \max_\pi V_\pi(s)$, for all $s \in \mathcal{S}$
- $Q^*(s,a) \triangleq \max_\pi Q_\pi(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}$

Learning $\boldsymbol{\pi^*(a|s)}$ through $\boldsymbol{V^*(s)}$ and $\boldsymbol{Q^*(s,a)}$ is related to the topic of **value-based RL**

Learning an explicit $\boldsymbol{\pi^*(a|s)}$ directly is related to the topic of **policy-based RL**

or both

RL Vocabulary

States: $s \in \mathcal{S}$

Actions: $a \in \mathcal{A}$

Policy: $\pi(a|s) \in [0, 1]$

Rewards: $r(s, a)$

Dynamics: $p(s'|s, a) \in [0, 1]$

Return: R_t

Value functions: $V_\pi(s), Q_\pi(s, a), V^*(s), Q^*(s, a)$

(Expected Return)

Questions

- Can you calculate $Q(s, a)$ for each state-action pairs w.r.t the uniform random policy in the chain env?
- What can be the optimal policy for the chain env with T=10?
- What will happened if we increase T? or change γ?