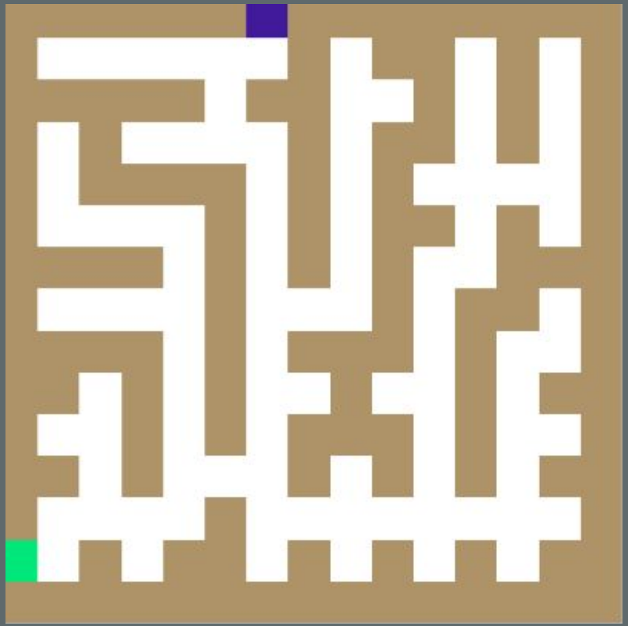
Maze Solver with Reinforcement Learning

Grigor Bezirganyan
Henrik Sergoyan

What is the project about?

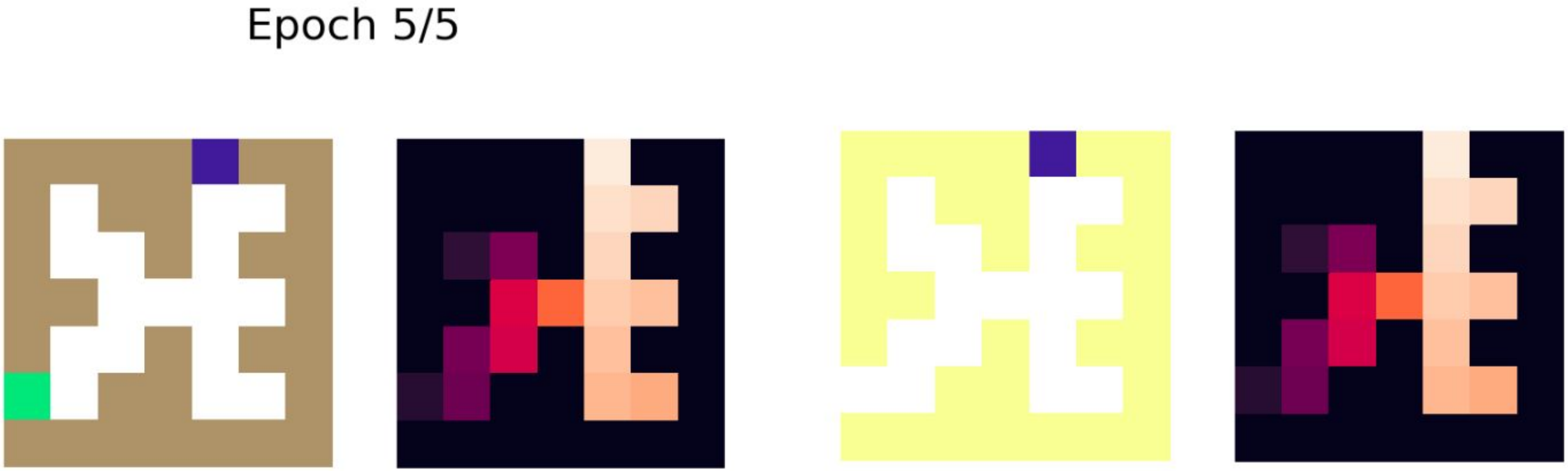


Steps

Background information

1. Creating the Environment
2. Learning Process
3. Developing the Optimal Policy
4. Solve Theeeeeeee Maaaaaze!

Creating the Environment



Learning Process

- Markov Property

$$p(\mathbf{s}_{t+1}|\mathbf{s}_0, \mathbf{s}_1, \ldots, \mathbf{s}_t, \mathbf{a}_t) = p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$$

- Reinforcement Learning

$$\mathbb{E}_{\tau_\pi}\left[\sum_{t=0}^{T} r(\mathbf{s}_t, \mathbf{a}_t)\right] \to \max_\pi$$

$$\mathbb{E}_{\tau_\pi}\left[\sum_{t=0}^{T} \gamma^t r(\mathbf{s}_t, \mathbf{a}_t)\right] \to \max_\pi, \quad 0 \leq \gamma \leq 1$$

discount factor

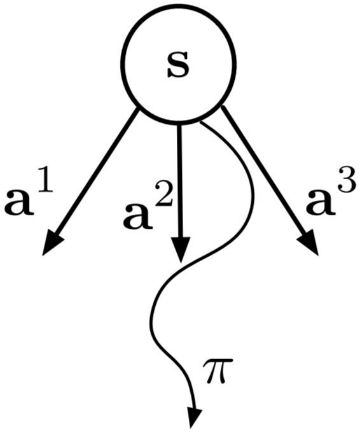
Value Function

- Value function shows how good it is to act in accordance to policy π starting from some state $\mathbf{s}_t$

$$V^\pi(\mathbf{s}_t) = \mathbb{E}_{\mathbf{a}_t, \mathbf{s}_{t+1}, \mathbf{a}_{t+1}, \ldots}\left[r(\mathbf{s}_t, \mathbf{a}_t) + \gamma r(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}) + \ldots\right]$$

- Optimal value function shows the maximum amount of reward we can get starting from some state $\mathbf{s}_t$

$$V^*(\mathbf{s}_t) = \max_\pi V^\pi(\mathbf{s}_t), \quad \forall \mathbf{s}_t \in \mathcal{S}$$

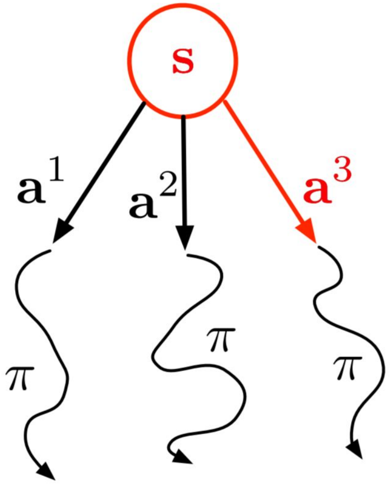


Q-function

- Q-function shows how good it is to act in accordance to policy π starting from some state $\mathbf{s}_t$ after taking action $\mathbf{a}_t$

$$Q^{\pi}(\mathbf{s}_t, \mathbf{a}_t) = \mathbb{E}_{\mathbf{s}_{t+1}, \mathbf{a}_{t+1}, \ldots}\left[r(\mathbf{s}_t, \mathbf{a}_t) + \gamma r(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}) + \ldots\right]$$

- Optimal Q-function shows the maximum amount of reward we can get starting from some state $\mathbf{s}_t$ after taking action $\mathbf{a}_t$



$$Q^{*}(\mathbf{s}_t, \mathbf{a}_t) = \max_{\pi} Q^{\pi}(\mathbf{s}_t, \mathbf{a}_t), \quad \forall(\mathbf{s}_t, \mathbf{a}_t)$$

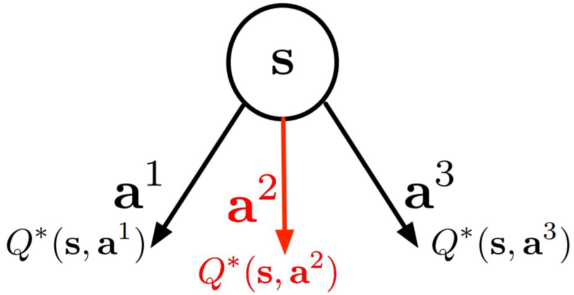
Optimal Policy

Suppose we know the optimal Q-function.

$$Q^*(\mathbf{s}_t, \mathbf{a}_t) = \max_{\pi} Q^{\pi}(\mathbf{s}_t, \mathbf{a}_t), \quad \forall (\mathbf{s}_t, \mathbf{a}_t)$$

What will be the optimal policy?

$$\pi^*(\mathbf{a}_t | \mathbf{s}_t) = \begin{cases} 1, & \mathbf{a}_t = \arg\max_{\mathbf{a}_t \in \mathcal{A}} Q^*(\mathbf{s}_t, \mathbf{a}_t) \\ 0, & \text{otherwise} \end{cases}$$



Thank You