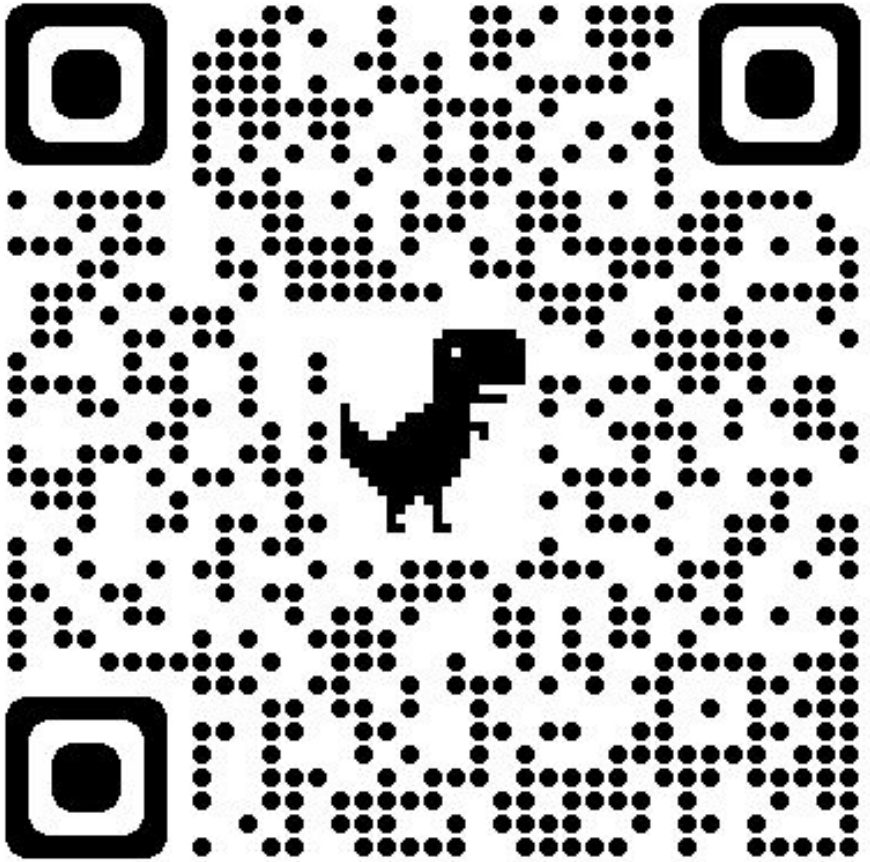
Discussion 13: On and Off Policy Learning



What we have learned so far

Solving RL problems

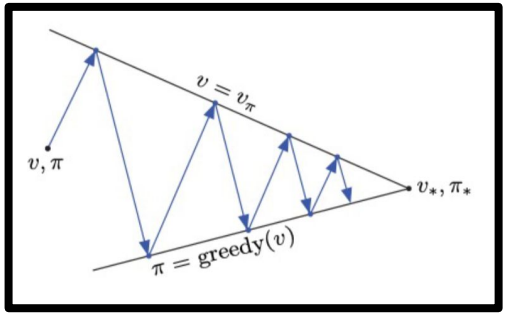
Solving RL problems with the techniques seen so far is an iterative process with two steps

Policy Evaluation

Compute v(s),q(s,a)

Policy Improvement

Compute π(a|s)



Use / Environment	Prediction	Control
Full knowledge	Dynamic programming	Dynamic programming
	Iterative policy evaluation	Policy iteration
Model free – on policy	Monte Carlo evaluation ✓ First visit MC ✓ Every-visit MC	Monte Carlo evaluation + ε – greedy policy improvement
Model free – on policy	Temporal difference evaluation	SARSA evaluation + ε – greedy policy improvement
Model free- off policy		Q-learning (SARSAMAX)

Policy Evaluation

Dynamic Programming:

$$v_{k+1}(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \left(\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_k(s') \right)$$

TD Learning:

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

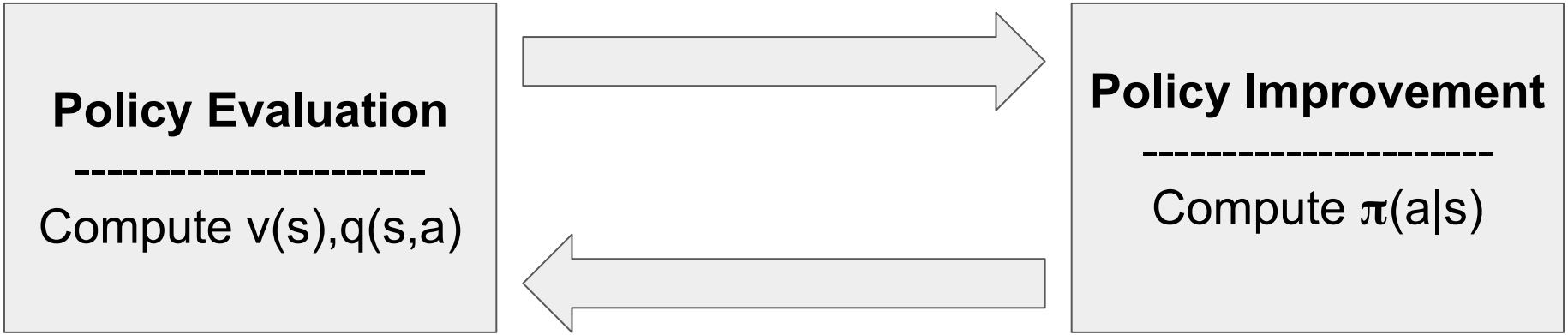
Q learning:

$$Q(S, A) \leftarrow Q(S, A) + \alpha \left(R + \gamma \max_{a'} Q(S', a') - Q(S, A) \right)$$

Why can on-policy methods fail?

Solving RL problems

Solving RL problems with the techniques seen so far is an iterative process with two steps



TD Learning:

Must keep visiting (S,A)

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

To answer this we must talk about ***The need for exploration?***

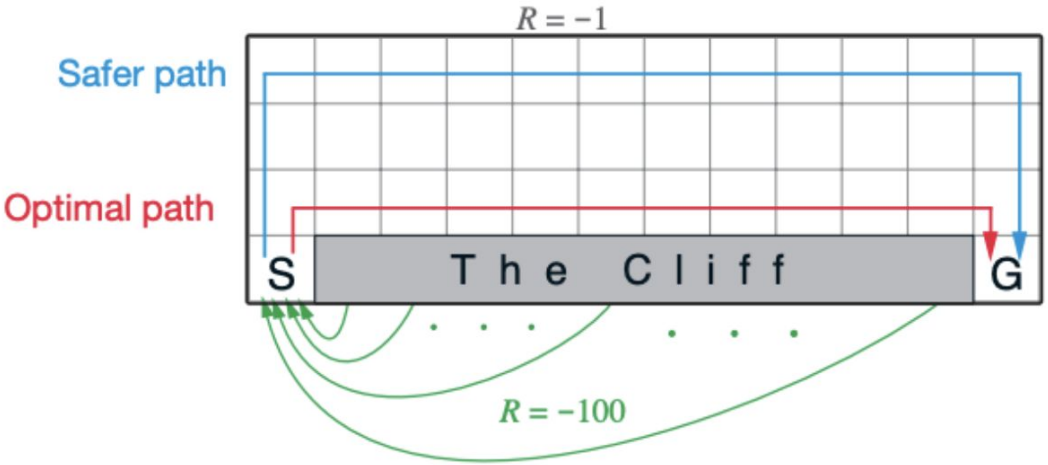
What is exploration?

Why do we need to explore?

Why could we choose the greedy action for Dynamic programming, but we need epsilon greedy for TD and Q learning?

We need to explore!

But sometimes exploration puts you in a dangerous spot



“On Policy” TD learning

TD Learning:

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

Current actions and future actions are chosen according to the same policy

“Off Policy” Q learning

Q Learning:

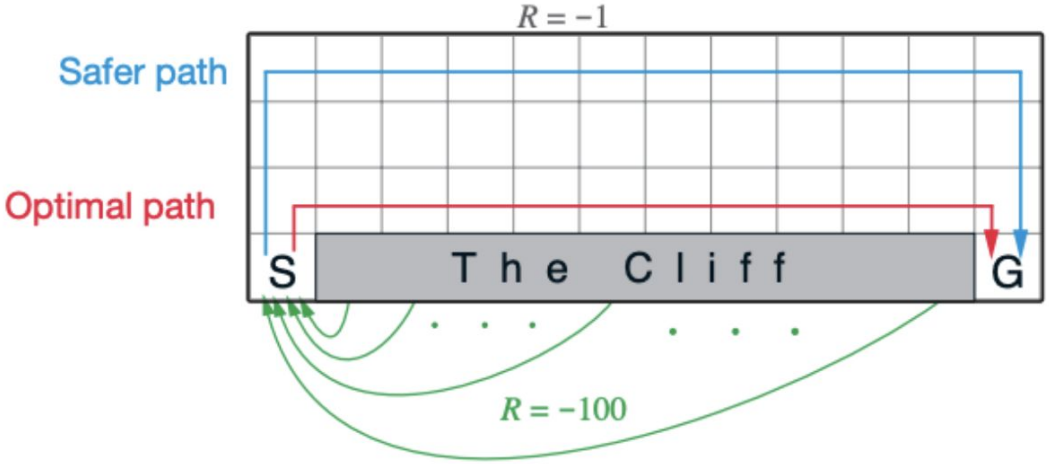
$$Q(S,A) \leftarrow Q(S,A) + \alpha \left(R + \gamma \max_{a'} Q(S', a') - Q(S,A) \right)$$

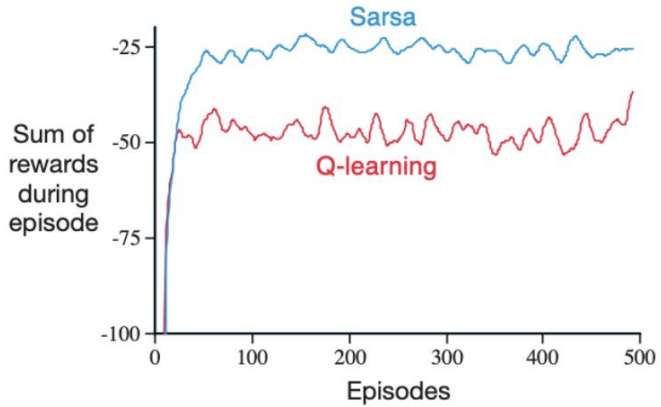
Current actions and future actions are chosen according to different policy

TD learning (SARSA) vs Q learning

TD learning converges to the optimal $Q_{\pi}(s,a)$ where π=epsilon-greedy

Q learning converges to $Q_{\pi}(s,a)$ where $\pi=\pi^*$ is optimal


R = −1
Safer path
Optimal path
S
The Cliff
G
R = −100


Sarsa
Q-learning
Sum of rewards during episode
-25
-50
-75
-100
0
100
200
300
400
500
Episodes

Coding tasks

Implement Q learning

Improve upon SARSA learning by tuning epsilon (change updateEpsilon())

Explore the effect of alpha on the convergence of both methods.