

# Lecture 5: Model-Free Control

David Silver

# Outline

- 1 Introduction
- 2 On-Policy Monte-Carlo Control
- 3 On-Policy Temporal-Difference Learning
- 4 Off-Policy Learning
- 5 Summary

# Model-Free Reinforcement Learning

- Last lecture:
  - **Model-free prediction**
  - *Estimate* the value function of an *unknown* MDP
- This lecture:
  - **Model-free control**
  - *Optimise* the value function of an *unknown* MDP

# Uses of Model-Free Control

Some example problems that can be modelled as MDPs

- Elevator
- Parallel Parking
- Ship Steering
- Bioreactor
- Helicopter
- Aeroplane Logistics
- Robocup Soccer
- Quake
- Portfolio management
- Protein Folding
- Robot walking
- Game of Go

For most of these problems, either:

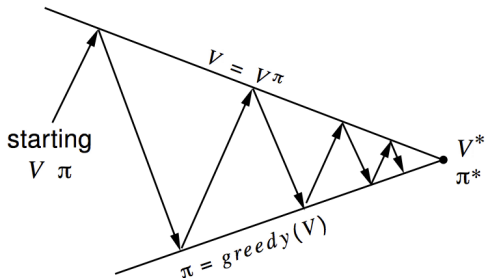
- MDP model is unknown, but experience can be sampled
- MDP model is known, but is too big to use, except by samples

**Model-free control** can solve these problems

# On and Off-Policy Learning

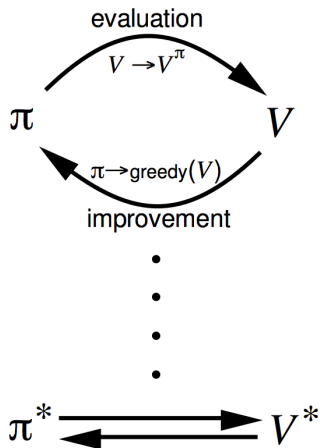
- **On-policy** learning
  - “Learn on the job”
  - Learn about policy  $\pi$  from experience sampled from  $\pi$
- **Off-policy** learning
  - “Look over someone’s shoulder”
  - Learn about policy  $\pi$  from experience sampled from  $\mu$

# Generalised Policy Iteration (Refresher)



**Policy evaluation** Estimate  $v_\pi$   
 e.g. Iterative policy evaluation

**Policy improvement** Generate  $\pi' \geq \pi$   
 e.g. Greedy policy improvement



# Generalised Policy Iteration With Monte-Carlo Evaluation



**Policy evaluation** Monte-Carlo policy evaluation,  $V = v_\pi$ ?

**Policy improvement** Greedy policy improvement?

# Model-Free Policy Iteration Using Action-Value Function

- Greedy policy improvement over  $V(s)$  requires model of MDP

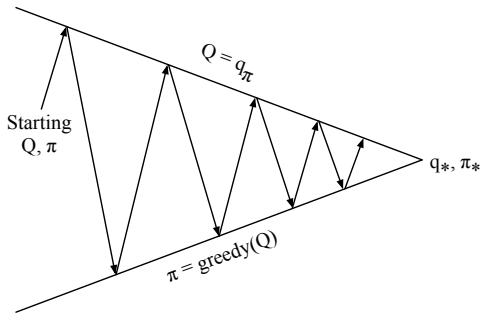
$$\pi'(s) = \operatorname{argmax}_{a \in \mathcal{A}} \mathcal{R}_s^a + \mathcal{P}_{ss'}^a V(s')$$

- Greedy policy improvement over  $Q(s, a)$  is model-free

$$\pi'(s) = \operatorname{argmax}_{a \in \mathcal{A}} Q(s, a)$$



# Generalised Policy Iteration with Action-Value Function



**Policy evaluation** Monte-Carlo policy evaluation,  $Q = q_\pi$

**Policy improvement** Greedy policy improvement?

## Example of Greedy Action Selection



"Behind one door is tenure - behind the other is flipping burgers at McDonald's."

- There are two doors in front of you.
- You open the left door and get reward 0  
 $V(\text{left}) = 0$
- You open the right door and get reward +1  
 $V(\text{right}) = +1$
- You open the right door and get reward +3  
 $V(\text{right}) = +2$
- You open the right door and get reward +2  
 $V(\text{right}) = +2$
- $\vdots$
- Are you sure you've chosen the best door?

# $\epsilon$ -Greedy Exploration

- Simplest idea for ensuring continual exploration
- All  $m$  actions are tried with non-zero probability
- With probability  $1 - \epsilon$  choose the greedy action
- With probability  $\epsilon$  choose an action at random

$$\pi(a|s) = \begin{cases} \epsilon/m + 1 - \epsilon & \text{if } a^* = \operatorname{argmax}_{a \in \mathcal{A}} Q(s, a) \\ \epsilon/m & \text{otherwise} \end{cases}$$

# $\epsilon$ -Greedy Policy Improvement

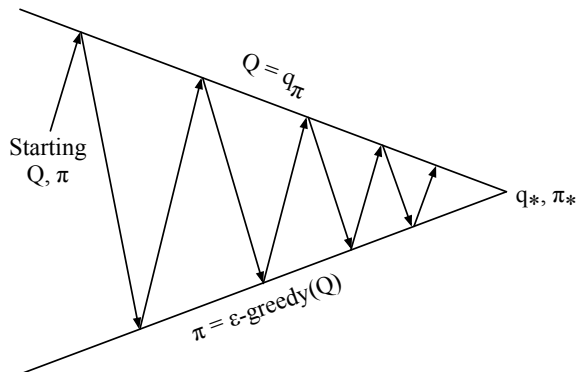
## Theorem

*For any  $\epsilon$ -greedy policy  $\pi$ , the  $\epsilon$ -greedy policy  $\pi'$  with respect to  $q_\pi$  is an improvement,  $v_{\pi'}(s) \geq v_\pi(s)$*

$$\begin{aligned} q_\pi(s, \pi'(s)) &= \sum_{a \in \mathcal{A}} \pi'(a|s) q_\pi(s, a) \\ &= \epsilon/m \sum_{a \in \mathcal{A}} q_\pi(s, a) + (1 - \epsilon) \max_{a \in \mathcal{A}} q_\pi(s, a) \\ &\geq \epsilon/m \sum_{a \in \mathcal{A}} q_\pi(s, a) + (1 - \epsilon) \sum_{a \in \mathcal{A}} \frac{\pi(a|s) - \epsilon/m}{1 - \epsilon} q_\pi(s, a) \\ &= \sum_{a \in \mathcal{A}} \pi(a|s) q_\pi(s, a) = v_\pi(s) \end{aligned}$$

Therefore from policy improvement theorem,  $v_{\pi'}(s) \geq v_\pi(s)$

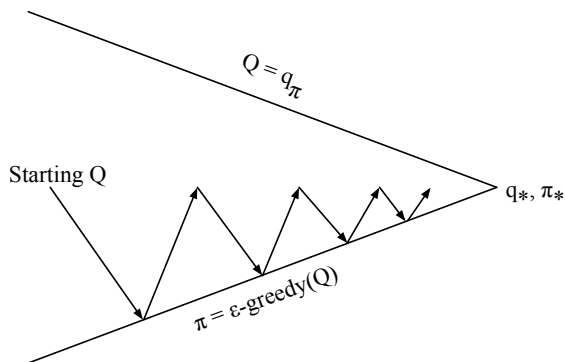
# Monte-Carlo Policy Iteration



Policy evaluation Monte-Carlo policy evaluation,  $Q = q_\pi$

Policy improvement  $\epsilon$ -greedy policy improvement

# Monte-Carlo Control



Every episode:

Policy evaluation Monte-Carlo policy evaluation,  $Q \approx q_\pi$

Policy improvement  $\epsilon$ -greedy policy improvement

# GLIE

## Definition

*Greedy in the Limit with Infinite Exploration (GLIE)*

- All state-action pairs are explored infinitely many times,

$$\lim_{k \rightarrow \infty} N_k(s, a) = \infty$$

- The policy converges on a greedy policy,

$$\lim_{k \rightarrow \infty} \pi_k(a|s) = \mathbf{1}(a = \operatorname{argmax}_{a' \in \mathcal{A}} Q_k(s, a'))$$

- For example,  $\epsilon$ -greedy is GLIE if  $\epsilon$  reduces to zero at  $\epsilon_k = \frac{1}{k}$

# GLIE Monte-Carlo Control

- Sample  $k$ th episode using  $\pi$ :  $\{S_1, A_1, R_2, \dots, S_T\} \sim \pi$
- For each state  $S_t$  and action  $A_t$  in the episode,

$$N(S_t, A_t) \leftarrow N(S_t, A_t) + 1$$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{1}{N(S_t, A_t)} (G_t - Q(S_t, A_t))$$

- Improve policy based on new action-value function

$$\epsilon \leftarrow 1/k$$

$$\pi \leftarrow \epsilon\text{-greedy}(Q)$$

## Theorem

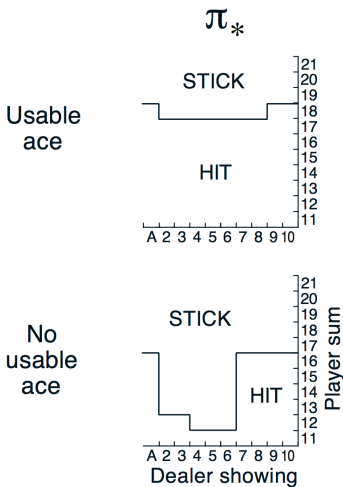
*GLIE Monte-Carlo control converges to the optimal action-value function,  $Q(s, a) \rightarrow q_*(s, a)$*



## Back to the Blackjack Example



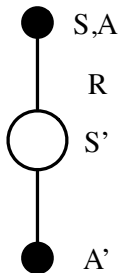
# Monte-Carlo Control in Blackjack



# MC vs. TD Control

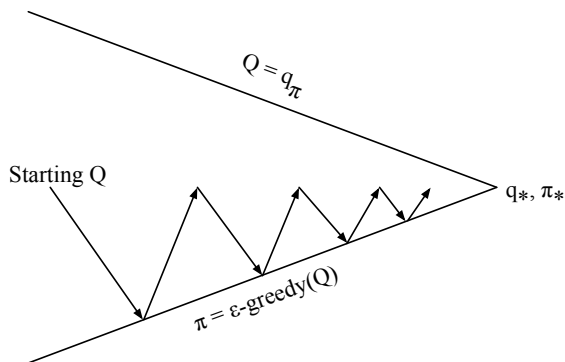
- Temporal-difference (TD) learning has several advantages over Monte-Carlo (MC)
  - Lower variance
  - Online
  - Incomplete sequences
- Natural idea: use TD instead of MC in our control loop
  - Apply TD to  $Q(S, A)$
  - Use  $\epsilon$ -greedy policy improvement
  - Update every time-step

# Updating Action-Value Functions with Sarsa



$$Q(S, A) \leftarrow Q(S, A) + \alpha (R + \gamma Q(S', A') - Q(S, A))$$

# On-Policy Control With Sarsa



Every **time-step**:

Policy evaluation **Sarsa**,  $Q \approx q_\pi$

Policy improvement  $\epsilon$ -greedy policy improvement

# Sarsa Algorithm for On-Policy Control

Initialize  $Q(s, a), \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$ , arbitrarily, and  $Q(\text{terminal-state}, \cdot) = 0$

Repeat (for each episode):

    Initialize  $S$

    Choose  $A$  from  $S$  using policy derived from  $Q$  (e.g.,  $\epsilon$ -greedy)

    Repeat (for each step of episode):

        Take action  $A$ , observe  $R, S'$

        Choose  $A'$  from  $S'$  using policy derived from  $Q$  (e.g.,  $\epsilon$ -greedy)

$Q(S, A) \leftarrow Q(S, A) + \alpha [R + \gamma Q(S', A') - Q(S, A)]$

$S \leftarrow S'; A \leftarrow A';$

    until  $S$  is terminal

# Convergence of Sarsa

## Theorem

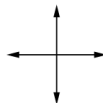
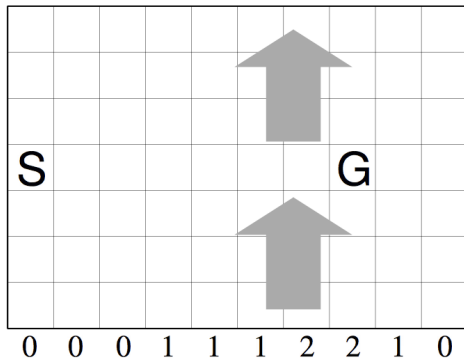
*Sarsa converges to the optimal action-value function,  $Q(s, a) \rightarrow q_*(s, a)$ , under the following conditions:*

- *GLIE sequence of policies  $\pi_t(a|s)$*
- *Robbins-Monro sequence of step-sizes  $\alpha_t$*

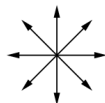
$$\sum_{t=1}^{\infty} \alpha_t = \infty$$

$$\sum_{t=1}^{\infty} \alpha_t^2 < \infty$$

# Windy Gridworld Example



standard  
moves

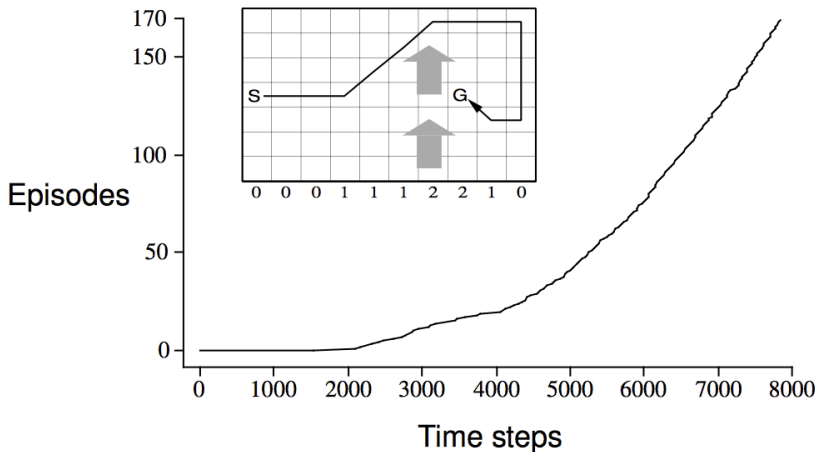


king's  
moves

- Reward = -1 per time-step until reaching goal
- Undiscounted



# Sarsa on the Windy Gridworld



## $n$ -Step Sarsa

- Consider the following  $n$ -step returns for  $n = 1, 2, \infty$ :

$$n = 1 \quad (\text{Sarsa}) \quad q_t^{(1)} = R_{t+1} + \gamma Q(S_{t+1})$$

$$n = 2 \quad q_t^{(2)} = R_{t+1} + \gamma R_{t+2} + \gamma^2 Q(S_{t+2})$$

$$\vdots$$
$$\vdots$$

$$n = \infty \quad (\text{MC}) \quad q_t^{(\infty)} = R_{t+1} + \gamma R_{t+2} + \dots + \gamma^{T-1} R_T$$

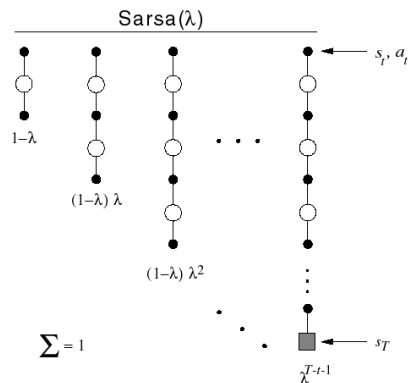
- Define the  $n$ -step Q-return

$$q_t^{(n)} = R_{t+1} + \gamma R_{t+2} + \dots + \gamma^{n-1} R_{t+n} + \gamma^n Q(S_{t+n})$$

- $n$ -step Sarsa updates  $Q(s, a)$  towards the  $n$ -step Q-return

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left( q_t^{(n)} - Q(S_t, A_t) \right)$$

# Forward View Sarsa( $\lambda$ )



- The  $q^\lambda$  return combines all  $n$ -step Q-returns  $q_t^{(n)}$
- Using weight  $(1 - \lambda)\lambda^{n-1}$

$$q_t^\lambda = (1 - \lambda) \sum_{n=1}^{\infty} \lambda^{n-1} q_t^{(n)}$$

- Forward-view Sarsa( $\lambda$ )

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left( q_t^\lambda - Q(S_t, A_t) \right)$$

# Backward View Sarsa( $\lambda$ )

- Just like TD( $\lambda$ ), we use **eligibility traces** in an online algorithm
- But Sarsa( $\lambda$ ) has one eligibility trace for each state-action pair

$$E_0(s, a) = 0$$

$$E_t(s, a) = \gamma\lambda E_{t-1}(s, a) + \mathbf{1}(S_t = s, A_t = a)$$

- $Q(s, a)$  is updated for every state  $s$  and action  $a$
- In proportion to TD-error  $\delta_t$  and eligibility trace  $E_t(s, a)$

$$\delta_t = R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)$$

$$Q(s, a) \leftarrow Q(s, a) + \alpha \delta_t E_t(s, a)$$

# Sarsa( $\lambda$ ) Algorithm

Initialize  $Q(s, a)$  arbitrarily, for all  $s \in \mathcal{S}, a \in \mathcal{A}(s)$

Repeat (for each episode):

$E(s, a) = 0$ , for all  $s \in \mathcal{S}, a \in \mathcal{A}(s)$

    Initialize  $S, A$

    Repeat (for each step of episode):

        Take action  $A$ , observe  $R, S'$

        Choose  $A'$  from  $S'$  using policy derived from  $Q$  (e.g.,  $\epsilon$ -greedy)

$\delta \leftarrow R + \gamma Q(S', A') - Q(S, A)$

$E(S, A) \leftarrow E(S, A) + \delta$

        For all  $s \in \mathcal{S}, a \in \mathcal{A}(s)$ :

$Q(s, a) \leftarrow Q(s, a) + \alpha \delta E(s, a)$

$E(s, a) \leftarrow \gamma \lambda E(s, a)$

$S \leftarrow S'; A \leftarrow A'$

    until  $S$  is terminal

A 10x10 grid with a black path. The path starts at (row, col) (0, 2), goes right to (0, 3), up to (1, 3), right to (1, 4), up to (2, 4), right to (2, 5), up to (3, 5), right to (3, 6), up to (4, 6), right to (4, 7), up to (5, 7), right to (5, 8), down to (6, 8), left to (6, 7), down to (7, 7), left to (7, 6), down to (8, 6), left to (8, 5), down to (9, 5), left to (9, 4), down to (10, 4), left to (10, 3), down to (11, 3), left to (11, 2), down to (12, 2), left to (12, 1), down to (13, 1), left to (13, 0), down to (14, 0), left to (14, 1), down to (15, 1), left to (15, 2), down to (16, 2), left to (16, 3), down to (17, 3), left to (17, 4), down to (18, 4), left to (18, 5), down to (19, 5), left to (19, 6), down to (20, 6), left to (20, 7), down to (21, 7), left to (21, 8), down to (22, 8), left to (22, 9), down to (23, 9), left to (23, 10), down to (24, 10), left to (24, 11), down to (25, 11), left to (25, 12), down to (26, 12), left to (26, 13), down to (27, 13), left to (27, 14), down to (28, 14), left to (28, 15), down to (29, 15), left to (29, 16), down to (30, 16), left to (30, 17), down to (31, 17), left to (31, 18), down to (32, 18), left to (32, 19), down to (33, 19), left to (33, 20), down to (34, 20), left to (34, 21), down to (35, 21), left to (35, 22), down to (36, 22), left to (36, 23), down to (37, 23), left to (37, 24), down to (38, 24), left to (38, 25), down to (39, 25), left to (39, 26), down to (40, 26), left to (40, 27), down to (41, 27), left to (41, 28), down to (42, 28), left to (42, 29), down to (43, 29), left to (43, 30), down to (44, 30), left to (44, 31), down to (45, 31), left to (45, 32), down to (46, 32), left to (46, 33), down to (47, 33), left to (47, 34), down to (48, 34), left to (48, 35), down to (49, 35), left to (49, 36), down to (50, 36), left to (50, 37), down to (51, 37), left to (51, 38), down to (52, 38), left to (52, 39), down to (53, 39), left to (53, 40), down to (54, 40), left to (54, 41), down to (55, 41), left to (55, 42), down to (56, 42), left to (56, 43), down to (57, 43), left to (57, 44), down to (58, 44), left to (58, 45), down to (59, 45), left to (59, 46), down to (60, 46), left to (60, 47), down to (61, 47), left to (61, 48), down to (62, 48), left to (62, 49), down to (63, 49), left to (63, 50), down to (64, 50), left to (64, 51), down to (65, 51), left to (65, 52), down to (66, 52), left to (66, 53), down to (67, 53), left to (67, 54), down to (68, 54), left to (68, 55), down to (69, 55), left to (69, 56), down to (70, 56), left to (70, 57), down to (71, 57), left to (71, 58), down to (72, 58), left to (72, 59), down to (73, 59), left to (73, 60), down to (74, 60), left to (74, 61), down to (75, 61), left to (75, 62), down to (76, 62), left to (76, 63), down to (77, 63), left to (77, 64), down to (78, 64), left to (78, 65), down to (79, 65), left to (79, 66), down to (80, 66), left to (80, 67), down to (81, 67), left to (81, 68), down to (82, 68), left to (82, 69), down to (83, 69), left to (83, 70), down to (84, 70), left to (84, 71), down to (85, 71), left to (85, 72), down to (86, 72), left to (86, 73), down to (87, 73), left to (87, 74), down to (88, 74), left to (88, 75), down to (89, 75), left to (89, 76), down to (90, 76), left to (90, 77), down to (91, 77), left to (91, 78), down to (92, 78), left to (92, 79), down to (93, 79), left to (93, 80), down to (94, 80), left to (94, 81), down to (95, 81), left to (95, 82), down to (96, 82), left to (96, 83), down to (97, 83), left to (97, 84), down to (98, 84), left to (98, 85), down to (99, 85), left to (99, 86), down to (100, 86), left to (100, 87), down to (101, 87), left to (101, 88), down to (102, 88), left to (102, 89), down to (103, 89), left to (103, 90), down to (104, 90), left to (104, 91), down to (105, 91), left to (105, 92), down to (106, 92), left to (106, 93), down to (107, 93), left to (107, 94), down to (108, 94), left to (108, 95), down to (109, 95), left to (109, 96), down to (110, 96), left to (110, 97), down to (111, 97), left to (111, 98), down to (112, 98), left to (112, 99), down to (113, 99), left to (113, 100), down to (114, 100), left to (114, 101), down to (115, 101), left to (115, 102), down to (116, 102), left to (116, 103), down to (117, 103), left to (117, 104), down to (118, 104), left to (118, 105), down to (119, 105), left to (119, 106), down to (120, 106), left to (120, 107), down to (121, 107), left to (121, 108), down to (122, 108), left to (122, 109), down to (123, 109), left to (123, 110), down to (124, 110), left to (124, 111), down to (125, 111), left to (125, 112), down to (126, 112), left to (126, 113), down to (127, 113), left to (127, 114), down to (128, 114), left to (128, 115), down to (129, 115), left to (129, 116), down to (130, 116), left to (130, 117), down to (131, 117), left to (131, 118), down to (132, 118), left to (132, 119), down to (133, 119), left to (133, 120), down to (134, 120), left to (134, 121), down to (135, 121), left to (135, 122), down to (136, 122), left to (136, 123), down to (137, 123), left to (137, 124), down to (138, 124), left to (138, 125), down to (139, 125), left to (139, 126), down to (140, 126), left to (140, 127), down to (141, 127), left to (141, 128), down to (142, 128), left to (142, 129), down to (143, 129), left to (143, 130), down to (144, 130), left to (144, 131), down to (145, 131), left to (145, 132), down to (146, 132), left to (146, 133), down to (147, 133), left to (147, 134), down to (148, 134), left to (148, 135), down to (149, 135), left to (149, 136), down to (150, 136), left to (150, 137), down to (151, 137), left to (151, 138), down to (152, 138), left to (152, 139), down to (153, 139), left to (153, 140), down to (154, 140), left to (154, 141), down to (155, 141), left to (155, 142), down to (156, 142), left to (156, 143), down to (157, 143), left to (157, 144), down to (158, 144), left to (158, 145), down to (159, 145), left to (159, 146), down to (160, 146), left to (160, 147), down to (161, 147), left to (161, 148), down to (162, 148), left to (162, 149), down to (163, 149), left to (163, 150), down to (164, 150), left to (164, 151), down to (165, 151), left to (165, 152), down to (166, 152), left to (166, 153), down to (167, 153), left to (167, 154), down to (168, 154), left to (168, 155), down to (169, 155), left to (169, 156), down to (170, 156), left to (170, 157), down to (171, 157), left to (171, 158), down to (172, 158), left to (172, 159), down to (173, 159), left to (173, 160), down to (174, 160), left to (174, 161), down to (175, 161), left to (175, 162), down to (176, 162), left to (176, 163), down to (177, 163), left to (177, 164), down to (178, 164), left to (178, 165), down to (179, 165), left to (179, 166), down to (180, 166), left to (180, 167), down to (181, 167), left to (181, 168), down to (182, 168), left to (182, 169), down to (183, 169), left to (183, 170), down to (184, 170), left to (184, 171), down to (185, 171), left to (185, 172), down to (186, 172), left to (186, 173), down to (187, 173), left to (187, 174), down to (188, 174), left to (188, 175), down to (189, 175

A 10x10 grid with a small black Christmas tree and a star on the 6th column, 4th row.

# Off-Policy Learning

- Evaluate target policy  $\pi(a|s)$  to compute  $v_\pi(s)$  or  $q_\pi(s, a)$
- While following behaviour policy  $\mu(a|s)$

$$\{S_1, A_1, R_2, \dots, S_T\} \sim \mu$$

- Why is this important?
- Learn from observing humans or other agents
- Re-use experience generated from old policies  $\pi_1, \pi_2, \dots, \pi_{t-1}$
- Learn about *optimal* policy while following *exploratory* policy
- Learn about *multiple* policies while following *one* policy

# Importance Sampling

- Estimate the expectation of a different distribution

$$\begin{aligned}\mathbb{E}_{X \sim P}[f(X)] &= \sum P(X)f(X) \\ &= \sum Q(X) \frac{P(X)}{Q(X)} f(X) \\ &= \mathbb{E}_{X \sim Q} \left[ \frac{P(X)}{Q(X)} f(X) \right]\end{aligned}$$



# Importance Sampling for Off-Policy Monte-Carlo

- Use returns generated from  $\mu$  to evaluate  $\pi$
- Weight return  $G_t$  according to similarity between policies
- Multiply importance sampling corrections along whole episode

$$G_t^{\pi/\mu} = \frac{\pi(A_t|S_t)}{\mu(A_t|S_t)} \frac{\pi(A_{t+1}|S_{t+1})}{\mu(A_{t+1}|S_{t+1})} \cdots \frac{\pi(A_T|S_T)}{\mu(A_T|S_T)} G_t$$

- Update value towards *corrected* return

$$V(S_t) \leftarrow V(S_t) + \alpha \left( G_t^{\pi/\mu} - V(S_t) \right)$$

- Cannot use if  $\mu$  is zero when  $\pi$  is non-zero
- Importance sampling can dramatically increase variance

# Importance Sampling for Off-Policy TD

- Use TD targets generated from  $\mu$  to evaluate  $\pi$
- Weight TD target  $R + \gamma V(S')$  by importance sampling
- Only need a single importance sampling correction

$$V(S_t) \leftarrow V(S_t) + \alpha \left( \frac{\pi(A_t|S_t)}{\mu(A_t|S_t)} (R_{t+1} + \gamma V(S_{t+1})) - V(S_t) \right)$$

- Much lower variance than Monte-Carlo importance sampling
- Policies only need to be similar over a single step

# Q-Learning

- We now consider off-policy learning of action-values  $Q(s, a)$
- **No** importance sampling is required
- Next action is chosen using behaviour policy  $A_{t+1} \sim \mu(\cdot|S_t)$
- But we consider alternative successor action  $A' \sim \pi(\cdot|S_t)$
- And update  $Q(S_t, A_t)$  towards value of alternative action

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha (R_{t+1} + \gamma Q(S_{t+1}, A') - Q(S_t, A_t))$$

# Off-Policy Control with Q-Learning

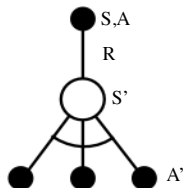
- We now allow both behaviour and target policies to **improve**
- The target policy  $\pi$  is **greedy** w.r.t.  $Q(s, a)$

$$\pi(S_{t+1}) = \operatorname{argmax}_{a'} Q(S_{t+1}, a')$$

- The behaviour policy  $\mu$  is e.g.  **$\epsilon$ -greedy** w.r.t.  $Q(s, a)$
- The Q-learning target then simplifies:

$$\begin{aligned} & R_{t+1} + \gamma Q(S_{t+1}, A') \\ &= R_{t+1} + \gamma Q(S_{t+1}, \operatorname{argmax}_{a'} Q(S_{t+1}, a')) \\ &= R_{t+1} + \max_{a'} \gamma Q(S_{t+1}, a') \end{aligned}$$

# Q-Learning Control Algorithm



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

## Theorem

*Q-learning control converges to the optimal action-value function,  $Q(s, a) \rightarrow q_*(s, a)$*

# Q-Learning Algorithm for Off-Policy Control

Initialize  $Q(s, a), \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$ , arbitrarily, and  $Q(\text{terminal-state}, \cdot) = 0$

Repeat (for each episode):

    Initialize  $S$

    Repeat (for each step of episode):

        Choose  $A$  from  $S$  using policy derived from  $Q$  (e.g.,  $\epsilon$ -greedy)

        Take action  $A$ , observe  $R, S'$

$Q(S, A) \leftarrow Q(S, A) + \alpha [R + \gamma \max_a Q(S', a) - Q(S, A)]$

$S \leftarrow S'$ ;

    until  $S$  is terminal

# Q-Learning Demo

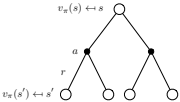
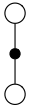
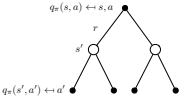

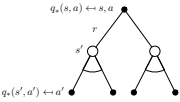
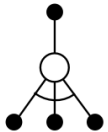
Q-Learning Demo

# Cliff Walking Example





# Relationship Between DP and TD

	<i>Full Backup (DP)</i>	<i>Sample Backup (TD)</i>
Bellman Expectation Equation for $v_{\pi}(s)$	 <p>Iterative Policy Evaluation</p>	 <p>TD Learning</p>
Bellman Expectation Equation for $q_{\pi}(s, a)$	 <p>Q-Policy Iteration</p>	 <p>Sarsa</p>
Bellman Optimality Equation for $q_{*}(s, a)$	 <p>Q-Value Iteration</p>	 <p>Q-Learning</p>

# Relationship Between DP and TD (2)

<i>Full Backup (DP)</i>	<i>Sample Backup (TD)</i>
Iterative Policy Evaluation $V(s) \leftarrow \mathbb{E}[R + \gamma V(S') \mid s]$	TD Learning $V(S) \stackrel{\alpha}{\leftarrow} R + \gamma V(S')$
Q-Policy Iteration $Q(s, a) \leftarrow \mathbb{E}[R + \gamma Q(S', A') \mid s, a]$	Sarsa $Q(S, A) \stackrel{\alpha}{\leftarrow} R + \gamma Q(S', A')$
Q-Value Iteration $Q(s, a) \leftarrow \mathbb{E}\left[R + \gamma \max_{a' \in \mathcal{A}} Q(S', a') \mid s, a\right]$	Q-Learning $Q(S, A) \stackrel{\alpha}{\leftarrow} R + \gamma \max_{a' \in \mathcal{A}} Q(S', a')$

where  $x \stackrel{\alpha}{\leftarrow} y \equiv x \leftarrow x + \alpha(y - x)$

# Questions?