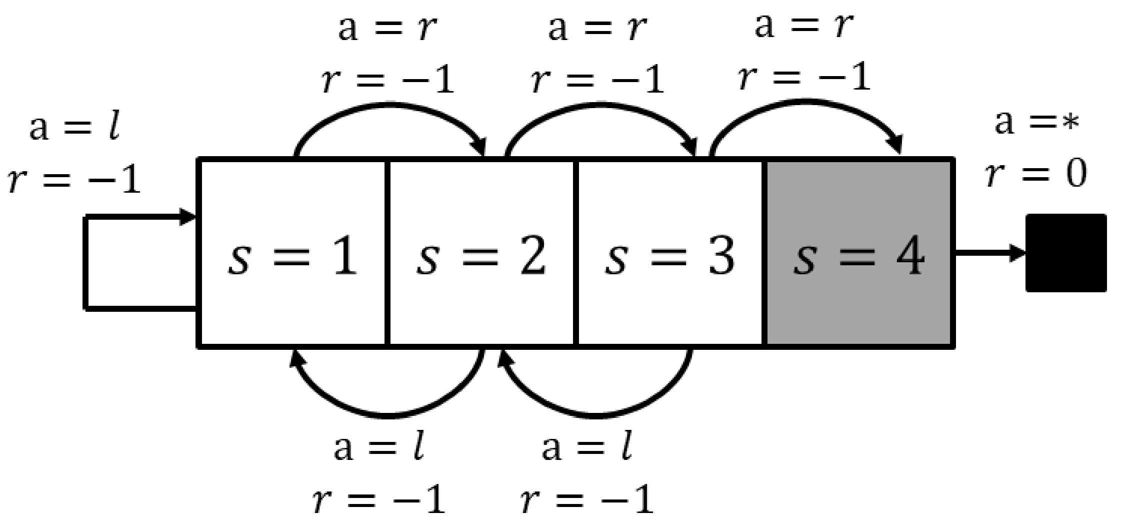
L7.2.X Linear Chain Example

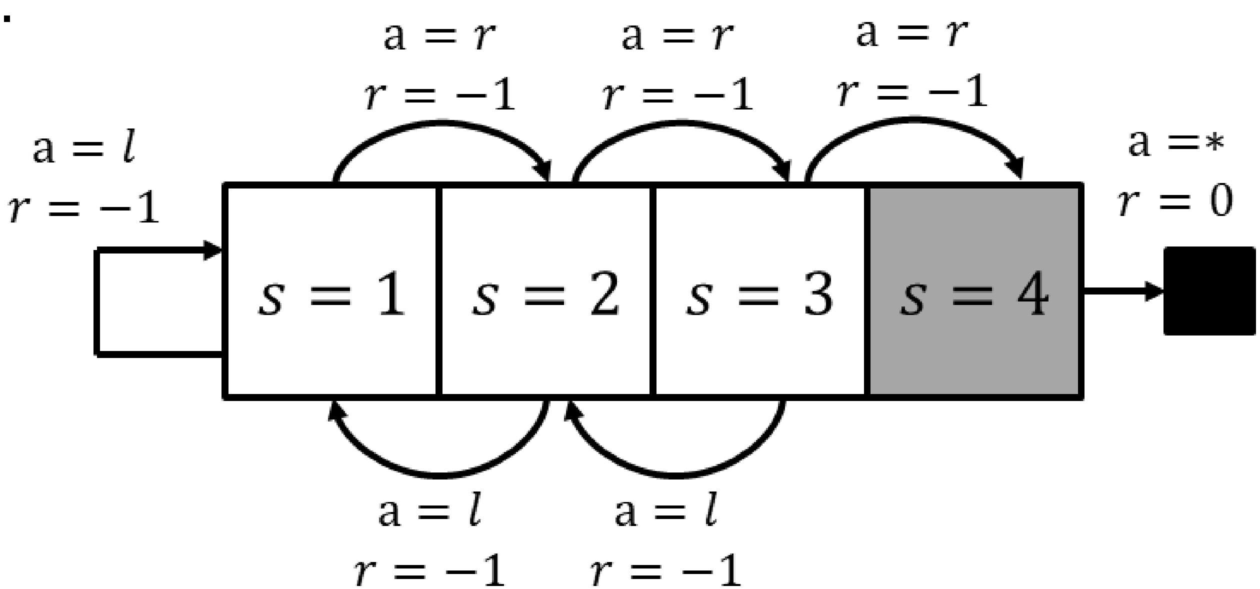
Zonghua Gu 2021



Acknowledgement: slides based on https://www.coursera.org/specializations/reinforcement-learning
And textbook by Sutton and Barto http://incompleteideas.net/book/the-book-2nd.html

Linear Chain Example

- Consider the following MDP. Environment is deterministic. In each state, there are two possible actions a∈{l,r}, where l corresponds to moving left, and r corresponds to moving right. Each movement incurs a reward of r=-1. State s=4 is the goal state: taking any action from s=4 results in reward of r=0 and ends the episode, hence $V(4) = 0, Q(4, a) = 0$ for any action a. Assume $\gamma = 1, \alpha = 1$. All value functions are initialized to 0.
- A. Use Policy Iteration, Value Iteration to derive optimal policy.



TD, Sarsa, QL

- B. Consider 8 consecutive episodes in the form of (s,a,r):

1. EP1: $(1, \mathrm{r}, -1), (2, \mathrm{r}, -1), (3, \mathrm{r}, -1), (4, \mathrm{r}, 0)$
2. EP2: $(1, \mathrm{r}, -1), (2, \mathrm{r}, -1), (3, \mathrm{r}, -1), (4, \mathrm{r}, 0)$
3. EP3: $(1, \mathrm{r}, -1), (2, \mathrm{r}, -1), (3, \mathrm{r}, -1), (4, \mathrm{r}, 0)$
4. EP4: $(3, \mathrm{l}, -1), (2, \mathrm{l}, -1), (1, \mathrm{l}, -1), (1, \mathrm{r}, -1), (2, \mathrm{r}, -1), (3, \mathrm{r}, -1), (4, \mathrm{r}, 0)$
5. EP5: $(3, \mathrm{l}, -1), (2, \mathrm{l}, -1), (1, \mathrm{l}, -1), (1, \mathrm{r}, -1), (2, \mathrm{r}, -1), (3, \mathrm{r}, -1), (4, \mathrm{r}, 0)$
6. EP6: $(3, \mathrm{l}, -1), (2, \mathrm{l}, -1), (1, \mathrm{l}, -1), (1, \mathrm{r}, -1), (2, \mathrm{r}, -1), (3, \mathrm{r}, -1), (4, \mathrm{r}, 0)$
7. EP7: $(3, \mathrm{l}, -1), (2, \mathrm{l}, -1), (1, \mathrm{l}, -1), (1, \mathrm{r}, -1), (2, \mathrm{r}, -1), (3, \mathrm{r}, -1), (4, \mathrm{r}, 0)$
8. EP8: $(3, \mathrm{l}, -1), (2, \mathrm{l}, -1), (1, \mathrm{l}, -1), (1, \mathrm{r}, -1), (2, \mathrm{r}, -1), (3, \mathrm{r}, -1), (4, \mathrm{r}, 0)$

- Derive the following (only show the changed parts):

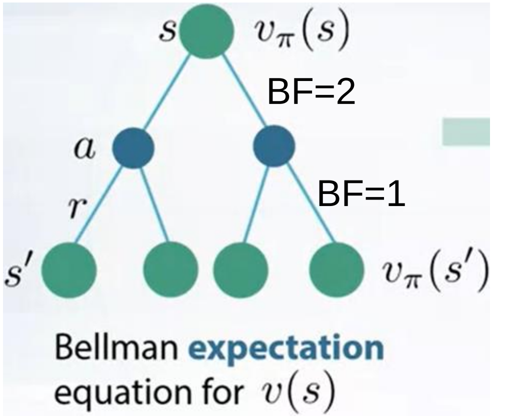
1. State value functions after TD learning.
2. State-action value functions (Q Value Functions) after Sarsa, and the resulting policy.
3. State-action value functions (Q Value Functions) after Q learning, and the resulting policy.

Recall: Simplified Bellman Equations for Deterministic Env

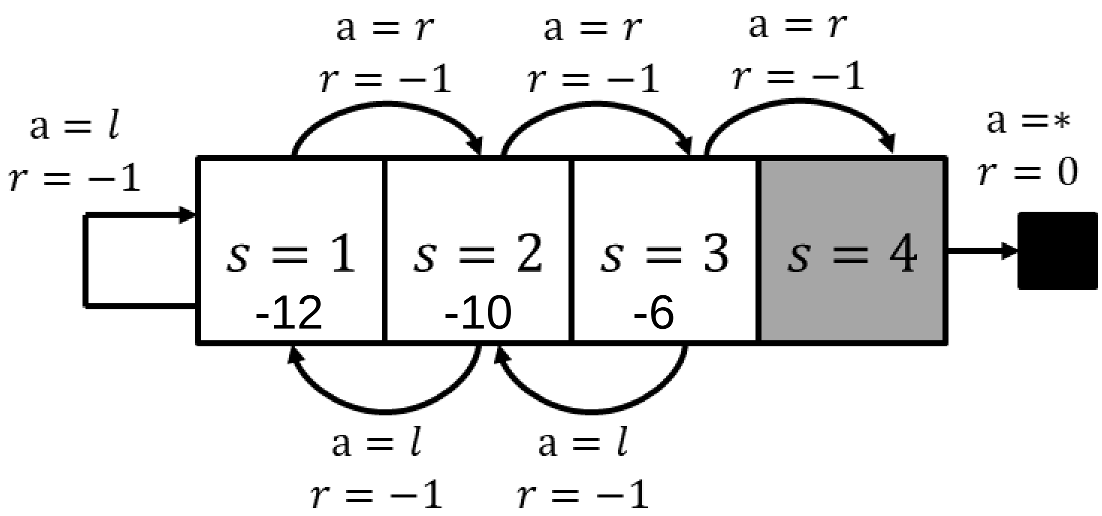
- Bellman Equations:
  - $v_\pi(s) = \sum_a \pi(a|s) q_\pi(s,a)$; $q_\pi(s,a) = \sum_{r,s'} p(r,s'|s,a)\,[r + \gamma v_\pi(s')]$
  - $v_*(s) = \max_a q_*(s,a)$; $q_*(s,a) = \sum_{r,s'} p(r,s'|s,a)\,[r + \gamma v_*(s')]$
- For Deterministic Env: there is only one possible (r, s') for a given (s, a) (we use R_s^a to emphasize that reward r is specific to this (s, a)):
  - $q_\pi(s,a) = R_s^a + \gamma v_\pi(s')$
  - $q_*(s,a) = R_s^a + \gamma v_*(s')$

Policy Iteration

1.1 Policy Evaluation of Random Policy

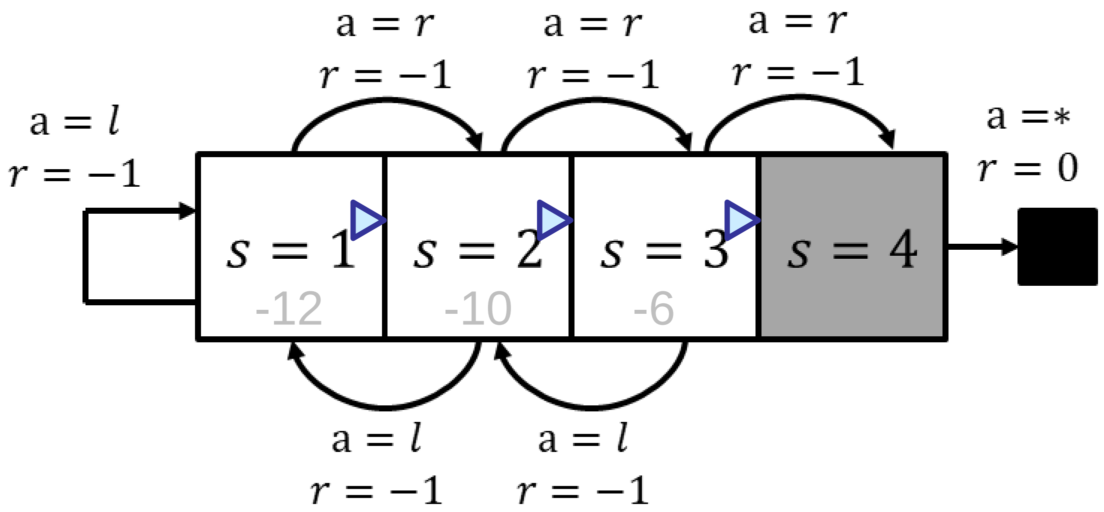


- Bellman Exp Equation: $v_\pi(s) = \sum_a \pi(a|s) q_\pi(s,a)$; $q_\pi(s,a) = R_s^a + \gamma v_\pi(s')$
- $Q_\pi(1,l) = -1 + v_\pi(1), Q_\pi(1,r) = -1 + v_\pi(2)$
- $v_\pi(1) = .5[Q_\pi(1,l) + Q_\pi(1,r)] = -1 + .5[v_\pi(1) + v_\pi(2)]$
- $Q_\pi(2,l) = -1 + v_\pi(1), Q_\pi(2,r) = -1 + v_\pi(3)$
- $v_\pi(2) = .5[Q_\pi(2,l) + Q_\pi(2,r)] = -1 + .5[v_\pi(1) + v_\pi(3)]$
- $Q_\pi(3,l) = -1 + v_\pi(2), Q_\pi(3,r) = -1 + v(4) = -1$
- $v_\pi(3) = .5[Q_\pi(3,l) + Q_\pi(3,r)] = -1 + .5\, v_\pi(2)$
- Solution: $v_\pi(1) = -12,\ v_\pi(2) = -10, v_\pi(3) = -6$

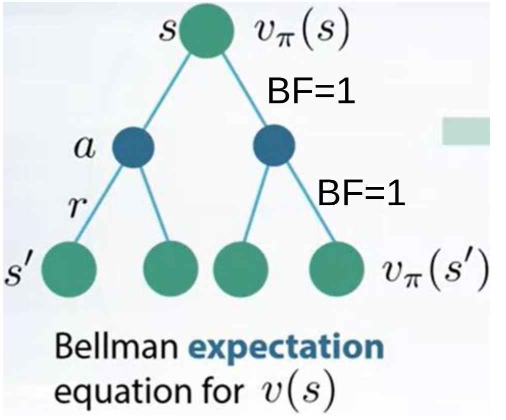


1.2 Policy Improvement

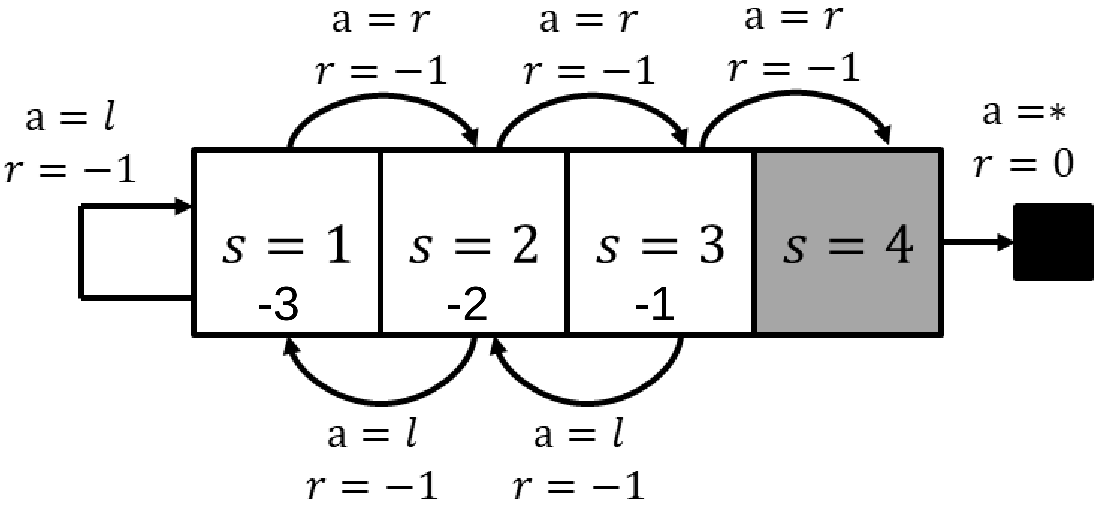
- Plug in values from PE to get new policy
- $Q_\pi(1,l) = -1 + v_\pi(1) = -13, Q_\pi(1,r) = -1 + v_\pi(2) = -11, \pi'(1) = \mathrm{argmax}_{\mathrm{a}}\big(Q_\pi(1,l), Q_\pi(1,r)\big) = r$
- $Q_\pi(2,l) = -1 + v_\pi(1) = -13, Q_\pi(2,r) = -1 + v_\pi(3) = -7, \pi'(2) = \mathrm{argmax}_{\mathrm{a}}\big(Q_\pi(2,l), Q_\pi(2,r)\big) = r$
- $Q_\pi(3,l) = -1 + v_\pi(2) = -11, Q_\pi(3,r) = -1, \pi'(3) = \mathrm{argmax}_{\mathrm{a}}\big(Q_\pi(3,l), Q_\pi(3,r)\big) = r$



2.1 Policy Evaluation of Det Policy

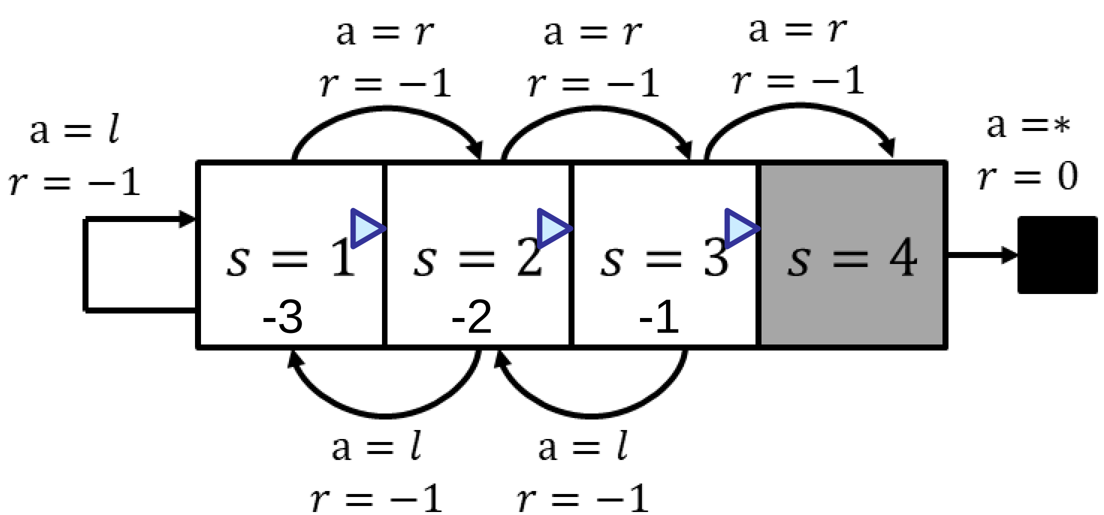


- $Q_\pi(1, r) = -1 + v_\pi(2)$
- $v_\pi(1) = 1Q_\pi(1, r) = -1 + v_\pi(2)$
- $Q_\pi(2, r) = -1 + v_\pi(3)$
- $v_\pi(2) = 1Q_\pi(2, r) = -1 + v_\pi(3)$
- $Q_\pi(3, r) = -1$
- $v_\pi(3) = 1Q_\pi(3, r) = -1$
- Solution: $v_\pi(1) = -3,\ v_\pi(2) = -2, v_\pi(3) = -1$



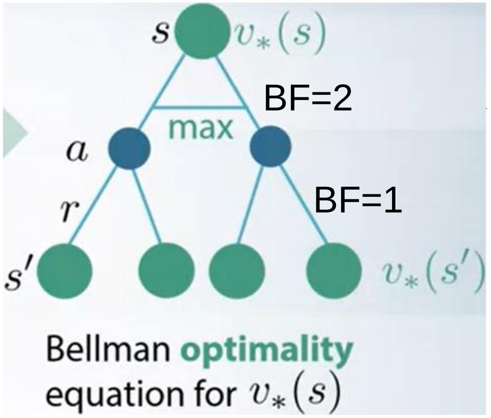
2.2 Policy Improvement

- Plug in values from PE to get new policy
- $Q_\pi(1,l) = -1 - 3 = -4$, $Q_\pi(1,r) = -1 - 2 = -3$, $\pi'(1) = \mathrm{argmax}_\mathrm{a}\big(Q_\pi(1,l), Q_\pi(1,r)\big) = r$
- $Q_\pi(2,l) = -1 - 3 = -4$, $Q_\pi(2,r) = -1 - 1 = -2$, $\pi'(2) = \mathrm{argmax}_\mathrm{a}\big(Q_\pi(2,l), Q_\pi(2,r)\big) = r$
- $Q_\pi(3,l) = -1 - 2 = -3$, $Q_\pi(3,r) = -1$, $\pi'(3) = \mathrm{argmax}_\mathrm{a}\big(Q_\pi(3,l), Q_\pi(3,r)\big) = r$
- Policy is now stable

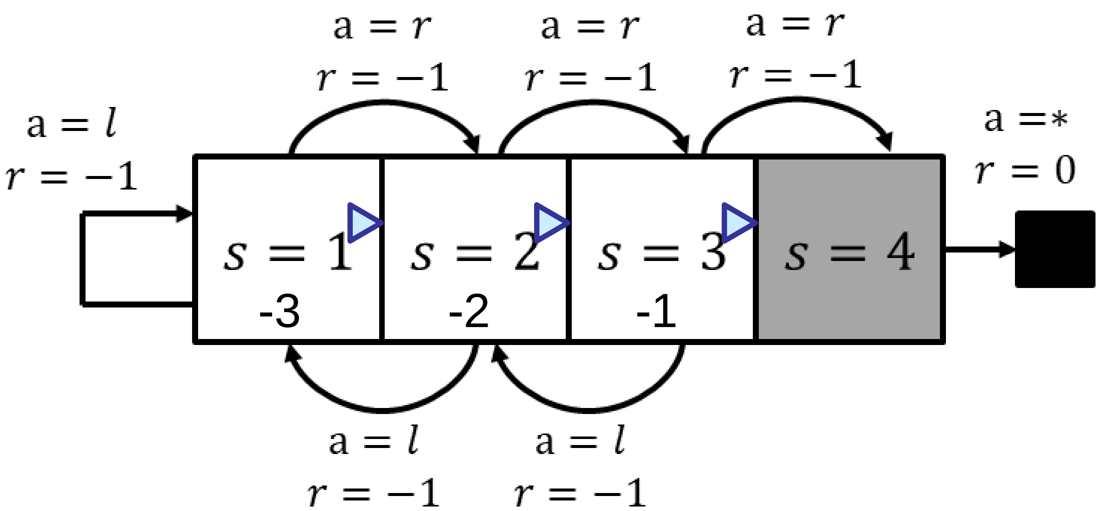


Value Iteration

Value Iteration



- Bellman Opt Equation: $v_*(s) = \max_a q_*(s,a); q_*(s,a) = R_S^a + \gamma v_*(s')$
- $Q_*(1,l) = -1 + v_*(1), Q_*(1,r) = -1 + v_*(2)$
- $v_*(1) = \max_a[Q_*(1,l), Q_*(1,r)] = \max_a[-1 + v_*(1), -1 + v_*(2)]$
- $Q_*(2,l) = -1 + v_*(1), Q_*(2,r) = -1 + v_*(3)$
- $v_*(2) = \max_a[Q_*(2,l), Q_*(2,r)] = \max_a[-1 + v_*(1), -1 + v_*(3)]$
- $Q_*(3,l) = -1 + v_*(2), Q_*(3,r) = -1 + v(4) = -1$
- $v_*(3) = \max_a[Q_*(3,l), Q_*(3,r)] = \max_a[-1 + v_*(2), -1 + v(4)] = \max_a[-1 + v_*(2), -1]$
- Solution: $v_*(1) = -3,\ v_*(2) = -2, v_*(3) = -1$
- W use Value Iteration to solve it. Table shows the iteration process until convergence
- Optimal policy: $\pi_*(1) = \underset{a\in(l,r)}{\operatorname{argmax}} Q_*(1,a) = r; \pi_*(2) = \underset{a\in(l,r)}{\operatorname{argmax}} Q_*(2,a) = r; \pi_*(3) = \underset{a\in(l,r)}{\operatorname{argmax}} Q_*(3,a) = r$



	$V_*(1)$	$V_*(2)$	$V_*(3)$
Init	0	0	0
Iter1	-1	-1	-1
Iter2	-2	-2	-1
Iter3	-3	-2	-1
Iter4	-3	-2	-1

TD Learning

EP1-3

- TD update equation: $V(S_t) \leftarrow V(S_t) + \alpha\big(R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\big) = R_{t+1} + V(S_{t+1})$
  - With $\gamma = 1, \alpha = 1$, each V(s) is completely replaced overwritten by the TD update
- $V(4) \equiv 0$. Initialize $V(1) = V(2) = V(3) = 0$,
- EP1: $(1, \mathrm{r}, -1), (2, \mathrm{r}, -1), (3, \mathrm{r}, -1), (4, \mathrm{r}, 0)$

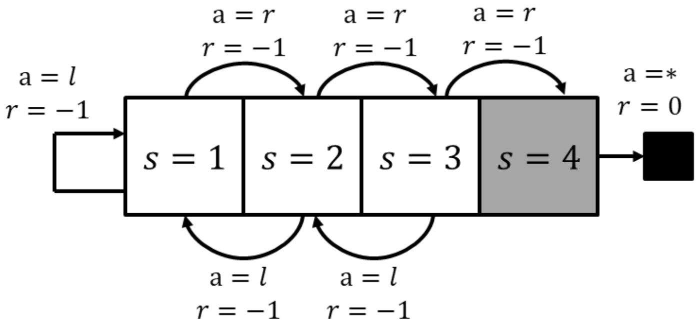
1. $V(1) \leftarrow -1 + V(2) = -1 + 0 = -1$
2. $V(2) \leftarrow -1 + V(3) = -1 + 0 = -1$
3. $V(3) \leftarrow -1 + V(4) = -1 + 0 = -1$

- EP2: $(1, \mathrm{r}, -1), (2, \mathrm{r}, -1), (3, \mathrm{r}, -1), (4, \mathrm{r}, 0)$

1. $V(1) \leftarrow -1 + V(2) = -1 - 1 = -2$
2. $V(2) \leftarrow -1 + V(3) = -1 - 1 = -2$
3. $V(3) \leftarrow -1 + V(4) = -1 + 0 = -1$

- EP3: $(1, \mathrm{r}, -1), (2, \mathrm{r}, -1), (3, \mathrm{r}, -1), (4, \mathrm{r}, 0)$

1. $V(1) \leftarrow -1 + V(2) = -1 - 2 = -3$
2. $V(2) \leftarrow -1 + V(3) = -1 - 1 = -2$
3. $V(3) \leftarrow -1 + V(4) = -1 + 0 = -1$



TD	$V(1)$	$V(2)$	$V(3)$
Init	0	0	0
After EP1	-1	-1	-1
After EP2	-2	-2	-1
After EP3	-3	-2	-1
After EP4	-5	-4	-1
After EP5	-7	-6	-1
After EP6	-9	-8	-1
After EP7	-11	-10	-1
After EP8	-13	-12	-1

- TD update equation: $V(S_t) \leftarrow R_{t+1} + V(S_{t+1})$

1. EP4:

   $(3, \mathrm{l}, -1), (2, \mathrm{l}, -1), (1, \mathrm{l}, -1), (1, \mathrm{r}, -1), (2, \mathrm{r}, -1), (3, \mathrm{r}, -1), (4, \mathrm{r}, 0)$

2. $V(3) \leftarrow -1 + V(2) = -1 - 2 = -3$
3. $V(2) \leftarrow -1 + V(1) = -1 - 3 = -4$
4. $V(1) \leftarrow -1 + V(1) = -1 - 3 = -4$
5. $V(1) \leftarrow -1 + V(2) = -1 - 4 = -5$
6. $V(2) \leftarrow -1 + V(3) = -1 - 3 = -4$
7. $V(3) \leftarrow -1 + V(4) = -1 + 0 = -1$

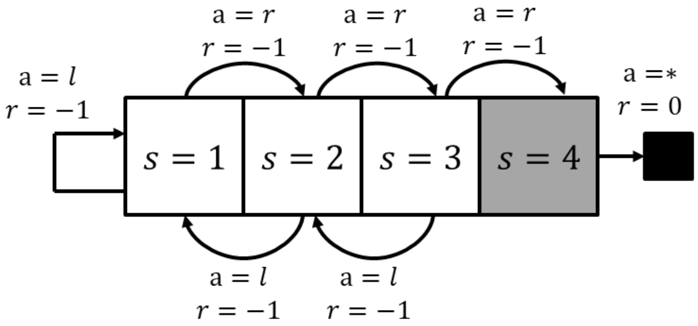
- EP5:

  $(3, \mathrm{l}, -1), (2, \mathrm{l}, -1), (1, \mathrm{l}, -1), (1, \mathrm{r}, -1), (2, \mathrm{r}, -1), (3, \mathrm{r}, -1), (4, \mathrm{r}, 0)$

1. $V(3) \leftarrow -1 + V(2) = -1 - 4 = -5$
2. $V(2) \leftarrow -1 + V(1) = -1 - 5 = -6$
3. $V(1) \leftarrow -1 + V(1) = -1 - 5 = -6$
4. $V(1) \leftarrow -1 + V(2) = -1 - 6 = -7$
5. $V(2) \leftarrow -1 + V(3) = -1 - 5 = -6$
6. $V(3) \leftarrow -1 + V(4) = -1 + 0 = -1$

- EP6-8 omitted

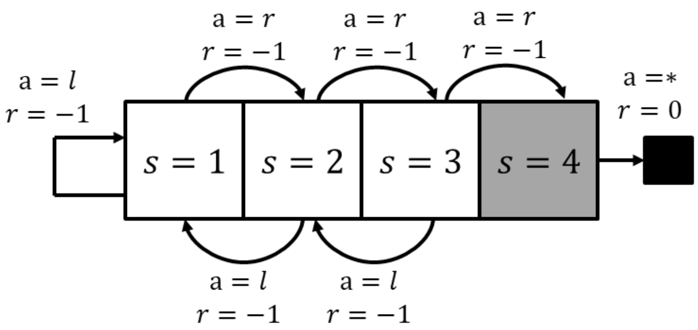
EP4-8



TD	$V(1)$	$V(2)$	$V(3)$
Init	0	0	0
After EP1	-1	-1	-1
After EP2	-2	-2	-1
After EP3	-3	-2	-1
After EP4	-5	-4	-1
After EP5	-7	-6	-1
After EP6	-9	-8	-1
After EP7	-11	-10	-1
After EP8	-13	-12	-1

TD failed to converge

- TD failed to converge for this set of episodes. The sequence of TD updates cause all value functions to be increasingly negative.
  - For simplicity, consider the infinite sequence of $(2, \mathrm{l}, -1), (1, \mathrm{r}, -1), (2, \mathrm{l}, -1)$ ...
  - The sequence of TD updates: $V(2) = -1 + V(1), V(1) = -1 + V(2)$, ... So $V(1)$ and $V(2)$ bootstrap off each other and both go to $-\infty$.
  - An analogy is that two students $V(1)$ and $V(2)$ are copying from each other, but they never get any true reward feedback from the teacher ($V(4) = 0$)



TD	$V(1)$	$V(2)$	$V(3)$
Init	0	0	0
After EP1	-1	-1	-1
After EP2	-2	-2	-1
After EP3	-3	-2	-1
After EP4	-5	-4	-1
After EP5	-7	-6	-1
After EP6	-9	-8	-1
After EP7	-11	-10	-1
After EP8	-13	-12	-1

- $V(3)$ is bootstrapped off $V(2)$ when moving left, and is bootstrapped off $V(4) \equiv 0$ when moving right. Steps 1-5 form a bootstrap dependency cycle $V(3) \leftarrow V(2) \leftarrow V(1) \leftarrow V(2) \leftarrow V(3)$ that causes $V(1), V(2), V(3)$ to blow up. Even though $V(3)$ is updated to $V(3) = -1 + V(4) = -1$ when it moves right to state 4, the episode ends immediately afterwards, so $V(1)$ and $V(2)$ do not have a chance to bootstrap off the correct $V(3)$.
  1. $V(3) \leftarrow -1 + V(2) = -1 - 4 = -5$
  2. $V(2) \leftarrow -1 + V(1) = -1 - 5 = -6$
  3. $V(1) \leftarrow -1 + V(1) = -1 - 5 = -6$
  4. $V(1) \leftarrow -1 + V(2) = -1 - 6 = -7$
  5. $V(2) \leftarrow -1 + V(3) = -1 - 5 = -6$
  6. $V(3) \leftarrow -1 + V(4) = -1 + 0 = -1$
- If the episode does not end immediately, but the agent moves left again, then $V(1)$ and $V(2)$ will have a chance to bootstrap off the new $V(3)$, and they may converge to the correct values.

Sarsa

- Sarsa update equation: $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\big(R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)\big) = R_{t+1} + Q(S_{t+1}, A_{t+1})$
  - With $\gamma = 1, \alpha = 1$, each Q(S, A) is completely replaced overwritten by the Sarsa update
- $Q(4, a) \equiv 0$. Initialize $Q(1,*) = Q(2,*) = Q(3,*) = 0$
- After EP1: $(1, \mathrm{r}, -1), (2, \mathrm{r}, -1), (3, \mathrm{r}, -1), (4, \mathrm{r}, 0)$

1. $Q(1, r) \leftarrow -1 + Q(2, r) = -1 + 0 = -1$
2. $Q(2, r) \leftarrow -1 + Q(3, r) = -1 + 0 = -1$
3. $Q(3, r) \leftarrow -1 + Q(4, r) = -1 + 0 = -1$

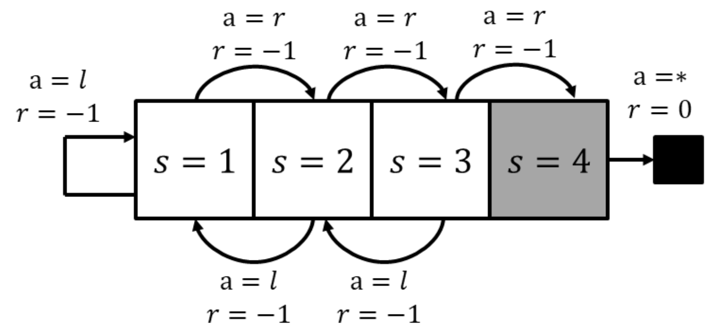
- After EP2: $(1, \mathrm{r}, -1), (2, \mathrm{r}, -1), (3, \mathrm{r}, -1), (4, \mathrm{r}, 0)$

1. $Q(1, r) \leftarrow -1 + Q(2, r) = -1 - 1 = -2$
2. $Q(2, r) \leftarrow -1 + Q(3, r) = -1 - 1 = -2$
3. $Q(3, r) \leftarrow -1 + Q(4, r) = -1 + 0 = -1$

- After EP3: $(1, \mathrm{r}, -1), (2, \mathrm{r}, -1), (3, \mathrm{r}, -1), (4, \mathrm{r}, 0)$

1. $Q(1, r) \leftarrow -1 + Q(2, r) = -1 - 2 = -3$
2. $Q(2, r) \leftarrow -1 + Q(3, r) = -1 - 1 = -2$
3. $Q(3, r) \leftarrow -1 + Q(4, r) = -1 + 0 = -1$

EP1-3



Sarsa	$Q(1, l)$	$\boldsymbol{Q(1, r)}$	$Q(2, l)$	$\boldsymbol{Q(2, r)}$	$Q(3, l)$	$\boldsymbol{Q(3, r)}$
Init	0	**0**	0	**0**	0	**0**
After EP1	0	**−1**	0	**−1**	0	**−1**
After EP2	0	**−2**	0	**−2**	0	**−1**
After EP3	0	**−3**	0	**−2**	0	**−1**
After EP4	−4	**−3**	−1	**−2**	−1	**−1**
After EP5	−4	**−3**	−5	**−2**	−2	**−1**
After EP6	−4	**−3**	−5	**−2**	−6	**−1**
After EP7	−4	**−3**	−5	**−2**	−6	**−1**
After EP8	−4	**−3**	−5	**−2**	−6	**−1**

EP4-8

- Sarsa update equation: $Q(S_t, A_t) \leftarrow R_{t+1} + Q(S_{t+1}, A_{t+1})$
- EP4: $(3, \mathrm{l}, -1), (2, \mathrm{l}, -1), (1, \mathrm{l}, -1), (1, \mathrm{r}, -1), (2, \mathrm{r}, -1), (3, \mathrm{r}, -1), (4, \mathrm{r}, 0)$

1. $Q(3, l) \leftarrow -1 + Q(2, l) = -1 + 0 = -1$
2. $Q(2, l) \leftarrow -1 + Q(1, l) = -1 + 0 = -1$
3. $Q(1, l) \leftarrow -1 + Q(1, r) = -1 - 3 = -4$
4. $Q(1, r) \leftarrow -1 + Q(2, r) = -1 - 2 = -3$
5. $Q(2, r) \leftarrow -1 + Q(3, r) = -1 - 1 = -2$
6. $Q(3, r) \leftarrow -1 + Q(4, r) = -1 + 0 = -1$

- EP5: $(3, \mathrm{l}, -1), (2, \mathrm{l}, -1), (1, \mathrm{l}, -1), (1, \mathrm{r}, -1), (2, \mathrm{r}, -1), (3, \mathrm{r}, -1), (4, \mathrm{r}, 0)$

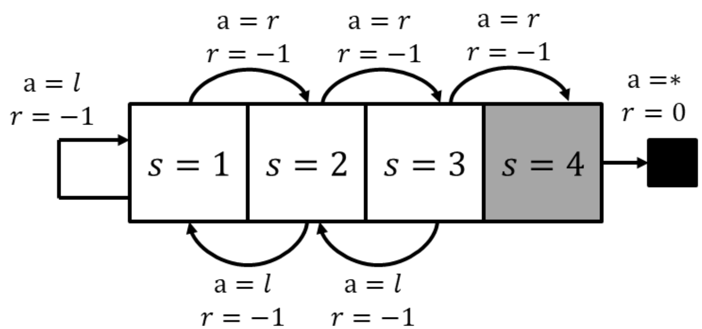
1. $Q(3, l) \leftarrow -1 + Q(2, l) = -1 - 1 = -2$
2. $Q(2, l) \leftarrow -1 + Q(1, l) = -1 - 4 = -5$
3. $Q(1, l) \leftarrow -1 + Q(1, r) = -1 - 3 = -4$
4. $Q(1, r) \leftarrow -1 + Q(2, r) = -1 - 2 = -3$
5. $Q(2, r) \leftarrow -1 + Q(3, r) = -1 - 1 = -2$
6. $Q(3, r) \leftarrow -1 + Q(4, r) = -1 + 0 = -1$

- EP6: $(3, \mathrm{l}, -1), (2, \mathrm{l}, -1), (1, \mathrm{l}, -1), (1, \mathrm{r}, -1), (2, \mathrm{r}, -1), (3, \mathrm{r}, -1), (4, \mathrm{r}, 0)$

1. $Q(3, l) \leftarrow -1 + Q(2, l) = -1 - 5 = -6$
2. $Q(2, l) \leftarrow -1 + Q(1, l) = -1 - 4 = -5$
3. $Q(1, l) \leftarrow -1 + Q(1, r) = -1 - 3 = -4$
4. $Q(1, r) \leftarrow -1 + Q(2, r) = -1 - 2 = -3$
5. $Q(2, r) \leftarrow -1 + Q(3, r) = -1 - 1 = -2$
6. $Q(3, r) \leftarrow -1 + Q(4, r) = -1 + 0 = -1$

- EP7: $(3, \mathrm{l}, -1), (2, \mathrm{l}, -1), (1, \mathrm{l}, -1), (1, \mathrm{r}, -1), (2, \mathrm{r}, -1), (3, \mathrm{r}, -1), (4, \mathrm{r}, 0)$

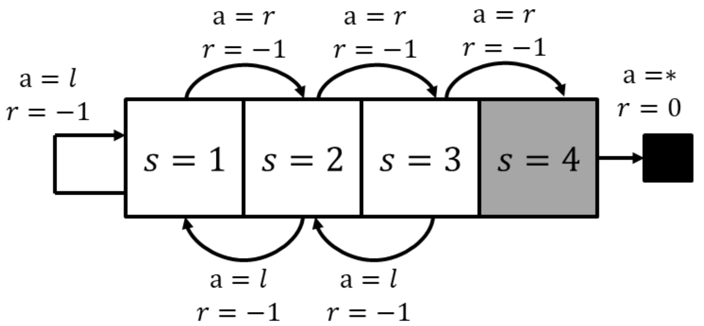
1. $Q(3, l) \leftarrow -1 + Q(2, l) = -1 - 5 = -6$
2. $Q(2, l) \leftarrow -1 + Q(1, l) = -1 - 4 = -5$
3. $Q(1, l) \leftarrow -1 + Q(1, r) = -1 - 3 = -4$
4. $Q(1, r) \leftarrow -1 + Q(2, r) = -1 - 2 = -3$
5. $Q(2, r) \leftarrow -1 + Q(3, r) = -1 - 1 = -2$
6. $Q(3, r) \leftarrow -1 + Q(4, r) = -1 + 0 = -1$ (Q values have converged, EP8 omitted)



Sarsa	$Q(1, l)$	$\boldsymbol{Q(1, r)}$	$Q(2, l)$	$\boldsymbol{Q(2, r)}$	$Q(3, l)$	$\boldsymbol{Q(3, r)}$
Init	0	$\mathbf{0}$	0	$\mathbf{0}$	0	$\mathbf{0}$
After EP1	0	$\mathbf{-1}$	0	$\mathbf{-1}$	0	$\mathbf{-1}$
After EP2	0	$\mathbf{-2}$	0	$\mathbf{-2}$	0	$\mathbf{-1}$
After EP3	0	$\mathbf{-3}$	0	$\mathbf{-2}$	0	$\mathbf{-1}$
After EP4	-4	$\mathbf{-3}$	-1	$\mathbf{-2}$	-1	$\mathbf{-1}$
After EP5	-4	$\mathbf{-3}$	-5	$\mathbf{-2}$	-2	$\mathbf{-1}$
After EP6	-4	$\mathbf{-3}$	-5	$\mathbf{-2}$	-6	$\mathbf{-1}$
After EP7	-4	$\mathbf{-3}$	-5	$\mathbf{-2}$	-6	$\mathbf{-1}$
After EP8	-4	$\mathbf{-3}$	-5	$\mathbf{-2}$	-6	$\mathbf{-1}$

Comments on Sarsa

- State-action value functions for moving right look reasonable: $Q(1, r) = -3, Q(2, r) = -2, Q(3, r) = -1$.
- State-action value functions for moving left look unreasonable: $Q(1, l) = -4, Q(2, l) = -5, Q(3, l) = -6$. This is because the only trajectory with move left actions are $3 \to 2 \to 1 \to 1 \to 2 \to 3 \to 4$, the Q values are updated based on only this episode (on-policy), i.e., from state 3 taking action left, it can only take the above trajectory, and reach the goal in 6 steps, hence $Q(3, l) = -6$. If we had collected more trajectories like $3 \to 2 \to 3 \to 4$, then Sarsa could learn the more accurate Q value $Q(3, l) = -1 + Q(2, r) = -3$.
- Even though the Q values for left actions are inaccurate, the greedy policy is still optimal since right action is always better than left:
- $\pi_*(1) = \text{argmax}_a\big(Q(1, l), Q(1, r)\big) = r$
- $\pi_*(2) = \text{argmax}_a\big(Q(2, l), Q(2, r)\big) = r$
- $\pi_*(3) = \text{argmax}_a\big(Q(3, l), Q(3, r)\big) = r$



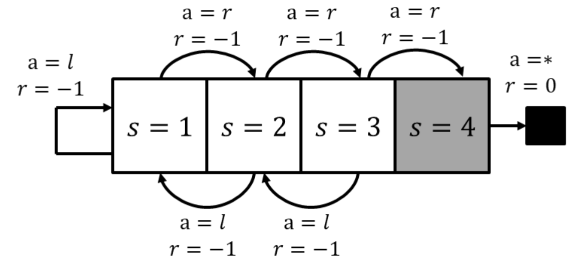
Sarsa	$Q(1, l)$	$\boldsymbol{Q(1, r)}$	$Q(2, l)$	$\boldsymbol{Q(2, r)}$	$Q(3, l)$	$\boldsymbol{Q(3, r)}$
Init	0	**0**	0	**0**	0	**0**
After EP1	0	**−1**	0	**−1**	0	**−1**
After EP2	0	**−2**	0	**−2**	0	**−1**
After EP3	0	**−3**	0	**−2**	0	**−1**
After EP4	−4	**−3**	−1	**−2**	−1	**−1**
After EP5	−4	**−3**	−5	**−2**	−2	**−1**
After EP6	−4	**−3**	−5	**−2**	−6	**−1**
After EP7	−4	**−3**	−5	**−2**	−6	**−1**
After EP8	−4	**−3**	−5	**−2**	−6	**−1**

Why Sarsa did not blow up

- TD: $V(s)$ is updated regardless if agent moves left or right. $V(3)$ is bootstrapped off $V(2)$ when moving left, and is bootstrapped off $V(4) \equiv 0$ when moving right. bootstrap dependency cycle $V(3) \leftarrow V(2) \leftarrow V(1) \leftarrow V(2) \leftarrow V(3)$ that causes $V(1), V(2), V(3)$ to blow up
- Sarsa: when agent moves left, $Q(s,l)$ is updated; when agent moves right, $Q(s,r)$ is updated. $Q(3,r)$ is always bootstrapped off $Q(4,r) \equiv 0$. So there is no bootstrap dependency cycle like TD. The linear dependency chain from $Q(4,r)$ to $Q(3,l)$ determines the stable values:

  1. $Q(3,l) \leftarrow -1 + Q(2,l) = -1 - 5 = -6$
  2. $Q(2,l) \leftarrow -1 + Q(1,l) = -1 - 4 = -5$
  3. $Q(1,l) \leftarrow -1 + Q(1,r) = -1 - 3 = -4$
  4. $Q(1,r) \leftarrow -1 + Q(2,r) = -1 - 2 = -3$
  5. $Q(2,r) \leftarrow -1 + Q(3,r) = -1 - 1 = -2$
  6. $Q(3,r) \leftarrow -1 + Q(4,r) = -1 + 0 = -1$

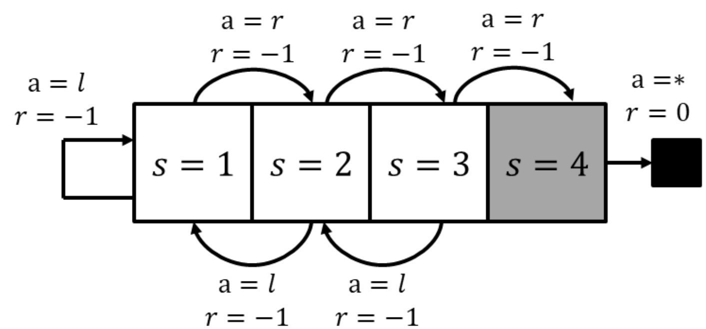
Sarsa	$Q(1,l)$	$\boldsymbol{Q(1,r)}$	$Q(2,l)$	$\boldsymbol{Q(2,r)}$	$Q(3,l)$	$\boldsymbol{Q(3,r)}$
Init	0	**0**	0	**0**	0	**0**
After EP1	0	**−1**	0	**−1**	0	**−1**
After EP2	0	**−2**	0	**−2**	0	**−1**
After EP3	0	**−3**	0	**−2**	0	**−1**
After EP4	−4	**−3**	−1	**−2**	−1	**−1**
After EP5	−4	**−3**	−5	**−2**	−2	**−1**
After EP6	−4	**−3**	−5	**−2**	−6	**−1**
After EP7	−4	**−3**	−5	**−2**	−6	**−1**
After EP8	−4	**−3**	−5	**−2**	−6	**−1**



Q Learning

EP1-3

- QL update equation: $(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left(R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t)\right) = R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a')$
  - With $\gamma = 1, \alpha = 1$, each Q(S, A) is completely replaced overwritten by the Q update



- After EP1: $(1, \mathrm{r}, -1), (2, \mathrm{r}, -1), (3, \mathrm{r}, -1), (4, \mathrm{r}, 0)$

1. $Q(1, r) \leftarrow -1 + \max_{a'} Q(2, a') = -1 + \max(0\,, 0) = -1$
2. $Q(2, r) \leftarrow -1 + \max_{a'} Q(3, a') = -1 + \max(0, 0) = -1$
3. $Q(3, r) \leftarrow -1 + \max_{a'} Q(4, a') = -1 + 0 = -1$

- After EP2: $(1, \mathrm{r}, -1), (2, \mathrm{r}, -1), (3, \mathrm{r}, -1), (4, \mathrm{r}, 0)$

1. $Q(1, r) \leftarrow -1 + \max_{a'} Q(2, a') = -1 + \max(-1,0) = -1$
2. $Q(2, r) \leftarrow -1 + \max_{a'} Q(3, a') = -1 + \max(-1,0) = -1$
3. $Q(3, r) \leftarrow -1 + \max_{a'} Q(4, a') = -1 + 0 = -1$

- After EP3: : $(1, \mathrm{r}, -1), (2, \mathrm{r}, -1), (3, \mathrm{r}, -1), (4, \mathrm{r}, 0)$

1. $Q(1, r) \leftarrow -1 + \max_{a'} Q(2, a') = -1 + \max(-1,0) = -1$
2. $Q(2, r) \leftarrow -1 + \max_{a'} Q(3, a') = -1 + \max(-1,0) = -1$
3. $Q(3, r) \leftarrow -1 + \max_{a'} Q(4, a') = -1 + 0 = -1$

QL	$Q(1, l)$	$\boldsymbol{Q(1, r)}$	$Q(2, l)$	$\boldsymbol{Q(2, r)}$	$Q(3, l)$	$\boldsymbol{Q(3, r)}$
Init	0	**0**	0	**0**	0	**0**
After EP1	0	**−1**	0	**−1**	0	**−1**
After EP2	0	**−1**	0	**−1**	0	**−1**
After EP3	0	**−1**	0	**−1**	0	**−1**
After EP4	−1	**−2**	−1	**−2**	−1	**−1**
After EP5	−2	−3	−2	**−2**	−2	**−1**
After EP6	−3	**−3**	−3	**−2**	−3	**−1**
After EP7	−4	**−3**	−4	**−2**	−3	**−1**
After EP8	−4	**−3**	−4	**−2**	−3	**−1**

EP4-6

- EP4: (3, l, −1), (2, l, −1), (1, l, −1), (1, r, −1), (2, r, −1), (3, r, −1), (4, r, 0)

1. $Q(3,l) \leftarrow -1 + \max_{a'} Q(2,a') = -1 + \max(0,-1) = -1$

2. $Q(2,l) \leftarrow -1 + \max_{a'} Q(1,a') = -1 + \max(0,-1) = -1$

3. $Q(1,l) \leftarrow -1 + \max_{a'} Q(1,a') = -1 + \max(0,-1) = -1$

4. $Q(1,r) \leftarrow -1 + \max_{a'} Q(2,a') = -1 + \max(-1,-1) = -2$

5. $Q(2,r) \leftarrow -1 + \max_{a'} Q(3,a') = -1 + \max(-1,-1) = -2$

6. $Q(3,r) \leftarrow -1 + \max_{a'} Q(4,a') = -1 + 0 = -1$

- EP5: (3, l, −1), (2, l, −1), (1, l, −1), (1, r, −1), (2, r, −1), (3, r, −1), (4, r, 0)

1. $Q(3,l) \leftarrow -1 + \max_{a'} Q(2,a') = -1 + \max(-1,-2) = -2$

2. $Q(2,l) \leftarrow -1 + \max_{a'} Q(1,a') = -1 + \max(-1,-2) = -2$

3. $Q(1,l) \leftarrow -1 + \max_{a'} Q(1,a') = -1 + \max(-1,-2) = -2$

4. $Q(1,r) \leftarrow -1 + \max_{a'} Q(2,a') = -1 + \max(-2,-2) = -3$

5. $Q(2,r) \leftarrow -1 + \max_{a'} Q(3,a') = -1 + \max(-2,-1) = -2$

6. $Q(3,r) \leftarrow -1 + \max_{a'} Q(4,a') = -1 + 0 = -1$

- EP6: (3, l, −1), (2, l, −1), (1, l, −1), (1, r, −1), (2, r, −1), (3, r, −1), (4, r, 0)

1. $Q(3,l) \leftarrow -1 + \max_{a'} Q(2,a') = -1 + \max(-2,-2) = -3$

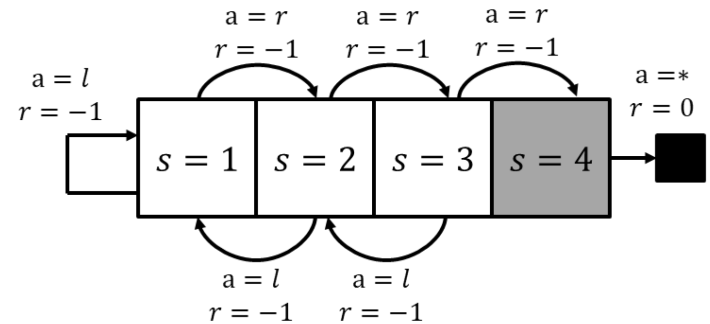
2. $Q(2,l) \leftarrow -1 + \max_{a'} Q(1,a') = -1 + \max(-2,-3) = -3$

3. $Q(1,l) \leftarrow -1 + \max_{a'} Q(1,a') = -1 + \max(-2,-3) = -3$

4. $Q(1,r) \leftarrow -1 + \max_{a'} Q(2,a') = -1 + \max(-3,-2) = -3$

5. $Q(2,r) \leftarrow -1 + \max_{a'} Q(3,a') = -1 + \max(-3,-1) = -2$

6. $Q(3,r) \leftarrow -1 + \max_{a'} Q(4,a') = -1 + 0 = -1$



QL	$Q(1,l)$	$\boldsymbol{Q(1,r)}$	$Q(2,l)$	$\boldsymbol{Q(2,r)}$	$Q(3,l)$	$\boldsymbol{Q(3,r)}$
Init	0	**0**	0	**0**	0	**0**
After EP1	0	**−1**	0	**−1**	0	**−1**
After EP2	0	**−1**	0	**−1**	0	**−1**
After EP3	0	**−1**	0	**−1**	0	**−1**
After EP4	−1	**−2**	−1	**−2**	−1	**−1**
After EP5	−2	**−3**	−2	**−2**	−2	**−1**
After EP6	−3	**−3**	−3	**−2**	−3	**−1**
After EP7	−4	**−3**	−4	**−2**	−3	**−1**
After EP8	−4	**−3**	−4	**−2**	−3	**−1**

EP7-8

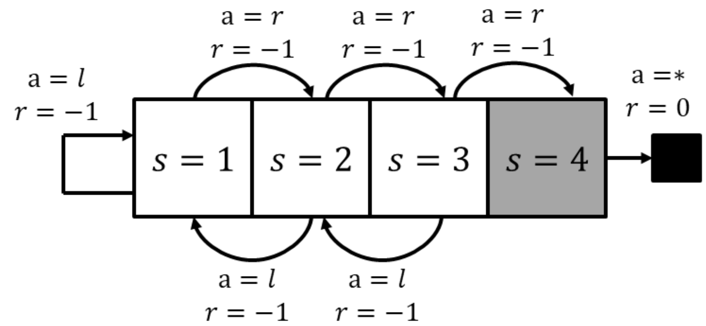
- EP7:
  (3, l, −1), (2, l, −1), (1, l, −1), (1, r, −1), (2, r, −1), (3, r, −1), (4, r, 0)

1. $Q(3,l) \leftarrow -1 + \max_{a'} Q(2,a') = -1 + \max(-3,-2) = -3$
2. $Q(2,l) \leftarrow -1 + \max_{a'} Q(1,a') = -1 + \max(-3,-3) = -4$
3. $Q(1,l) \leftarrow -1 + \max_{a'} Q(1,a') = -1 + \max(-3,-3) = -4$
4. $Q(1,r) \leftarrow -1 + \max_{a'} Q(2,a') = -1 + \max(-4,-2) = -3$
5. $Q(2,r) \leftarrow -1 + \max_{a'} Q(3,a') = -1 + \max(-3,-1) = -2$
6. $Q(3,r) \leftarrow -1 + \max_{a'} Q(4,a') = -1 + 0 = -1$

- EP8:
  (3, l, −1), (2, l, −1), (1, l, −1), (1, r, −1), (2, r, −1), (3, r, −1), (4, r, 0)

1. $Q(3,l) \leftarrow -1 + \max_{a'} Q(2,a') = -1 + \max(-4,-2) = -3$
2. $Q(2,l) \leftarrow -1 + \max_{a'} Q(1,a') = -1 + \max(-4,-3) = -4$
3. $Q(1,l) \leftarrow -1 + \max_{a'} Q(1,a') = -1 + \max(-4,-3) = -4$
4. $Q(1,r) \leftarrow -1 + \max_{a'} Q(2,a') = -1 + \max(-4,-2) = -3$
5. $Q(2,r) \leftarrow -1 + \max_{a'} Q(3,a') = -1 + \max(-3,-1) = -2$
6. $Q(3,r) \leftarrow -1 + \max_{a'} Q(4,a') = -1 + 0 = -1$

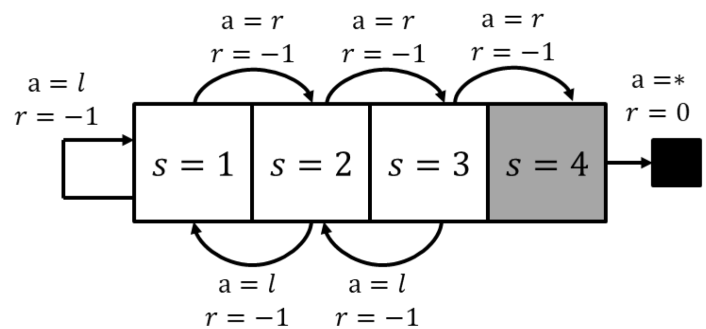
- Q values have converged



QL	$Q(1,l)$	$\boldsymbol{Q(1,r)}$	$Q(2,l)$	$\boldsymbol{Q(2,r)}$	$Q(3,l)$	$\boldsymbol{Q(3,r)}$
Init	0	**0**	0	**0**	0	**0**
After EP1	0	**−1**	0	**−1**	0	**−1**
After EP2	0	**−1**	0	**−1**	0	**−1**
After EP3	0	**−1**	0	**−1**	0	**−1**
After EP4	−1	**−2**	−1	**−2**	−1	**−1**
After EP5	−2	−3	−2	**−2**	−2	**−1**
After EP6	−3	**−3**	−3	**−2**	−3	**−1**
After EP7	−4	**−3**	−4	**−2**	−3	**−1**
After EP8	−4	**−3**	−4	**−2**	−3	**−1**

Comments on QL

- QL converges. All state-action value functions look reasonable:$Q(1,r)=-3, Q(2,r)=-2, Q(3,r)=-1$.
- $Q(1,l)=-4$: If agent moves left in state 1, it takes at most 4 steps to reach goal state 4: $1\rightarrow 1\rightarrow 2\rightarrow 3\rightarrow 4$.
- $Q(2,l)=-4$: If agent moves left in state 2, it takes at most 4 steps to reach goal state 4: $2\rightarrow 1\rightarrow 2\rightarrow 3\rightarrow 4$.
- $Q(3,l)=-3$: If agent moves left in state 3, it takes at most 3 steps to reach goal state 4: $3\rightarrow 2\rightarrow 3\rightarrow 4$.
- So QL is smarter than Sarsa: since it is off-policy, agent can learn the correct Q value functions that correspond to trajectories that it has never actually experienced, e.g., if If agent moves left in state 3, even though it never experienced the trajectory $3\rightarrow 2\rightarrow 3\rightarrow 4$, its Q values are updated so that the optimal policy follows that trajectory.
- Q values learned by QL are accurate, and the greedy policy is optimal:
- $\pi_*(1)=\mathrm{argmax}_{\mathrm{a}}\big(Q(1,l),Q(1,r)\big)=r$
- $\pi_*(2)=\mathrm{argmax}_{\mathrm{a}}\big(Q(2,l),Q(2,r)\big)=r$
- $\pi_*(3)=\mathrm{argmax}_{\mathrm{a}}\big(Q(3,l),Q(3,r)\big)=r$



QL	$Q(1,l)$	$\boldsymbol{Q(1,r)}$	$Q(2,l)$	$\boldsymbol{Q(2,r)}$	$Q(3,l)$	$\boldsymbol{Q(3,r)}$
Init	0	**0**	0	**0**	0	**0**
After EP1	0	**−1**	0	**−1**	0	**−1**
After EP2	0	**−1**	0	**−1**	0	**−1**
After EP3	0	**−1**	0	**−1**	0	**−1**
After EP4	−1	**−2**	−1	**−2**	−1	**−1**
After EP5	−2	**−3**	−2	**−2**	−2	**−1**
After EP6	−3	**−3**	−3	**−2**	−3	**−1**
After EP7	−4	**−3**	−4	**−2**	−3	**−1**
After EP8	−4	**−3**	−4	**−2**	−3	**−1**