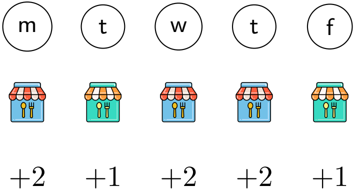
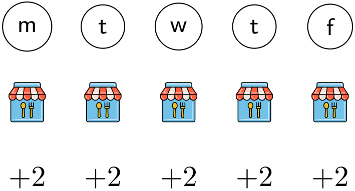
Exploration vs. exploitation

Real life example of greedy policy improvement issues



m	t	w	t	f
$+2$	$+1$	$+2$	$+2$	$+1$

Figure: Week 1



m	t	w	t	f
$+2$	$+2$	$+2$	$+2$	$+2$

Figure: Week 2

Real life example of greedy policy improvement issues

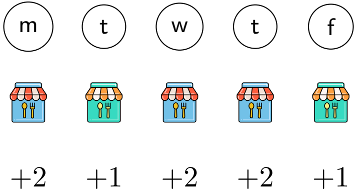


Figure: Week 1

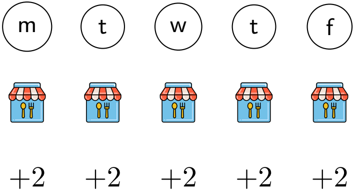
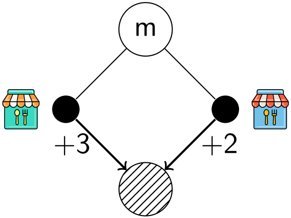


Figure: Week 2



Real life example of greedy policy improvement issues

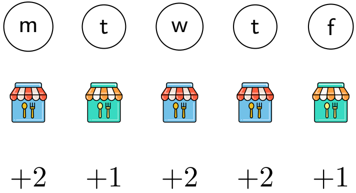


Figure: Week 1

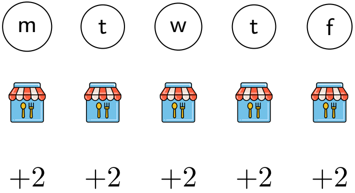


Figure: Week 2

- If random policy is too short, we don't see some state action pairs. We have no Q-value estimates for them.
  - $Q($ m , 🏪 $) = 3$

Real life example of greedy policy improvement issues

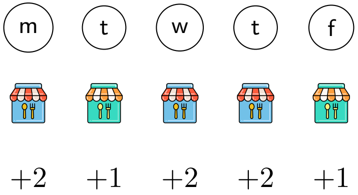


Figure: Week 1

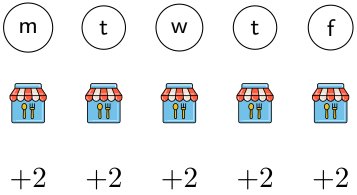


Figure: Week 2

- If random policy is too short, we don't see some state action pairs. We have no Q-value estimates for them.
    - $Q($ m , 🏪 $) = 3$

- Greedy policy will never encounter this state-action pair

Real life example of greedy policy improvement issues

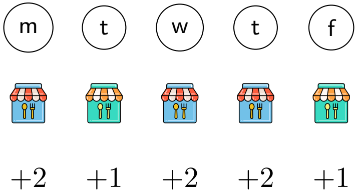


Figure: Week 1

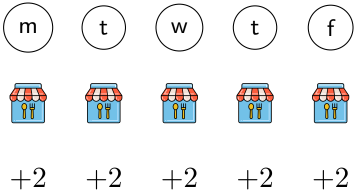


Figure: Week 2

- If random policy is too short, we don't see some state action pairs. We have no Q-value estimates for them.
  - $Q($ⓜ, 🏪$) = 3$

- Greedy policy will never encounter this state-action pair
- Even worse at Q-value discovery!

Summary: Exploration vs. exploitation

Issues

Summary: Exploration vs. exploitation

Issues

- If we don't **explore** (using random actions) enough, we don't see all state-action pairs. We don't know their Q-values.

Summary: Exploration vs. exploitation

Issues

- If we don't **explore** (using random actions) enough, we don't see all state-action pairs. We don't know their Q-values.
- If we can't estimate Q-values, we can't do greedy policy improvement.

Summary: Exploration vs. exploitation

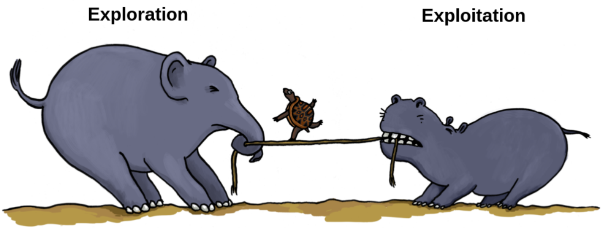
Issues

- If we don't **explore** (using random actions) enough, we don't see all state-action pairs. We don't know their Q-values.
- If we can't estimate Q-values, we can't do greedy policy improvement.
- If we don't **exploit** (policy improvement), we don't get closer to our goal of finding the optimal policy

Summary: Exploration vs. exploitation

Issues

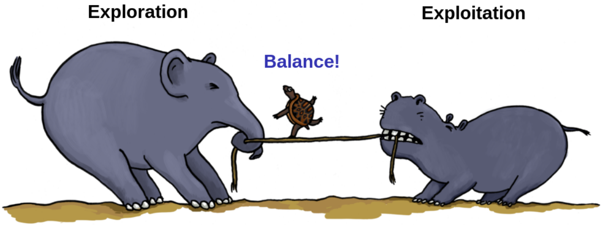
- If we don't **explore** (using random actions) enough, we don't see all state-action pairs. We don't know their Q-values.
- If we can't estimate Q-values, we can't do greedy policy improvement.
- If we don't **exploit** (policy improvement), we don't get closer to our goal of finding the optimal policy



Summary: Exploration vs. exploitation

Issues

- If we don't **explore** (using random actions) enough, we don't see all state-action pairs. We don't know their Q-values.
- If we can't estimate Q-values, we can't do greedy policy improvement.
- If we don't **exploit** (policy improvement), we don't get closer to our goal of finding the optimal policy



Summary: Exploration vs. exploitation

Issues

- If we don't **explore** (using random actions) enough, we don't see all state-action pairs. We don't know their Q-values.
- If we can't estimate Q-values, we can't do greedy policy improvement.
- If we don't **exploit** (policy improvement), we don't get closer to our goal of finding the optimal policy

Solution

- Need to balance exploration and exploitation in any RL problem

Summary: Exploration vs. exploitation

Issues

- If we don't **explore** (using random actions) enough, we don't see all state-action pairs. We don't know their Q-values.
- If we can't estimate Q-values, we can't do greedy policy improvement.
- If we don't **exploit** (policy improvement), we don't get closer to our goal of finding the optimal policy

Solution

- Need to balance exploration and exploitation in any RL problem
    - Greedy policy improvement is all exploitation and no exploration.

Summary: Exploration vs. exploitation

Issues

- If we don't **explore** (using random actions) enough, we don't see all state-action pairs. We don't know their Q-values.
- If we can't estimate Q-values, we can't do greedy policy improvement.
- If we don't **exploit** (policy improvement), we don't get closer to our goal of finding the optimal policy

Solution

- Need to balance exploration and exploitation in any RL problem
  - Greedy policy improvement is all exploitation and no exploration.
  - Next lesson: add an exploration component to greedy policy improvement.