UNIVERSITY OF LEEDS

Class: Machine Learning

Policy Iteration

Instructor: Matteo Leonetti

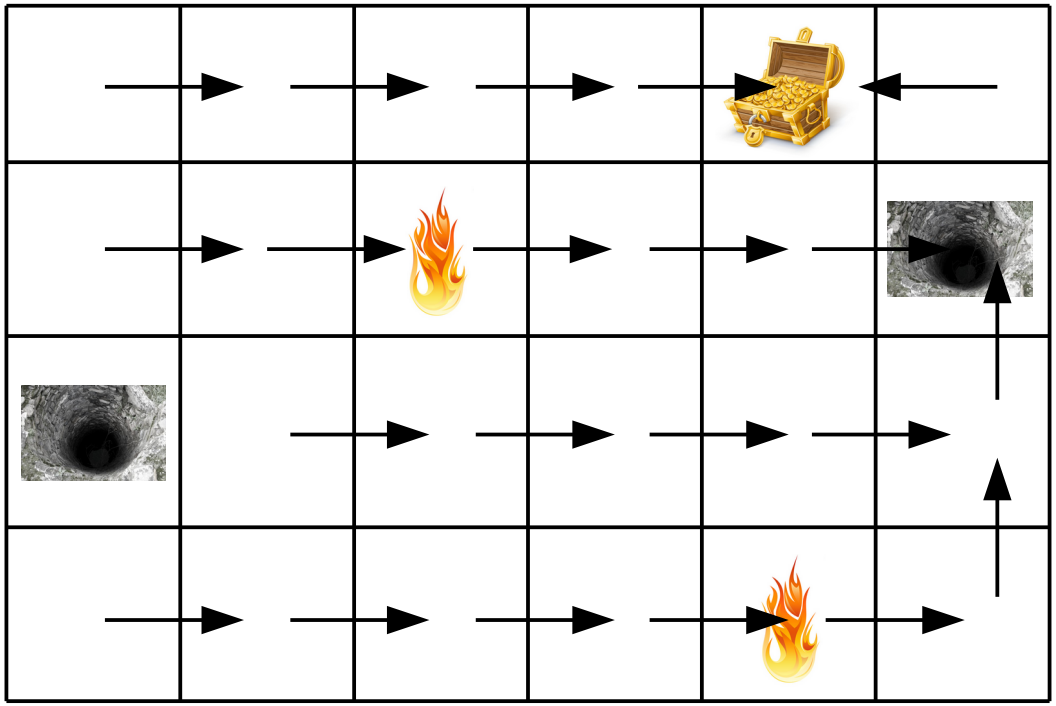
Learning outcomes

UNIVERSITY OF LEEDS

- Define and compute the value and action value function for a given policy
- Perform a step of generalized policy iteration

Example: policy

UNIVERSITY OF LEEDS



Is this a good policy? How do we establish the value of a policy?

Value

UNIVERSITY OF LEEDS

We define the *value* of a state under a given policy, as the expected return from that state while following the policy:

$$G_t \equiv \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} = R_{t+1} + \gamma G_{t+1}$$

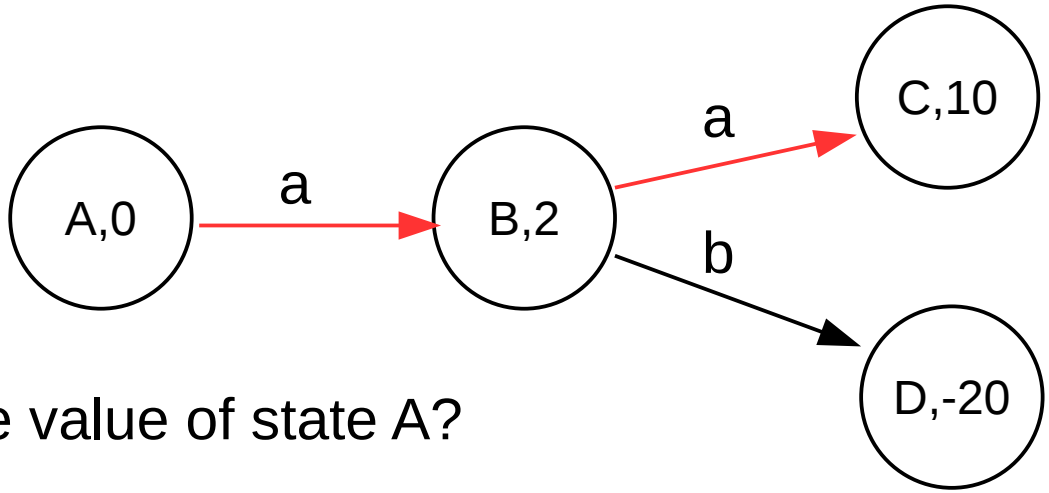
$$v_\pi(s) \equiv E_\pi[G_t | S_t = s] = E_\pi[R_{t+1}] + \gamma E_\pi[G_{t+1}]$$

$$= E_\pi[R_t] + \gamma v(s_{t+1})$$

This is called the *Bellman* equation

Understanding the value

UNIVERSITY OF LEEDS



What is the value of state A?

It depends on what the agent is going to do in B (the policy)

Assuming:

$\gamma = 0.5$

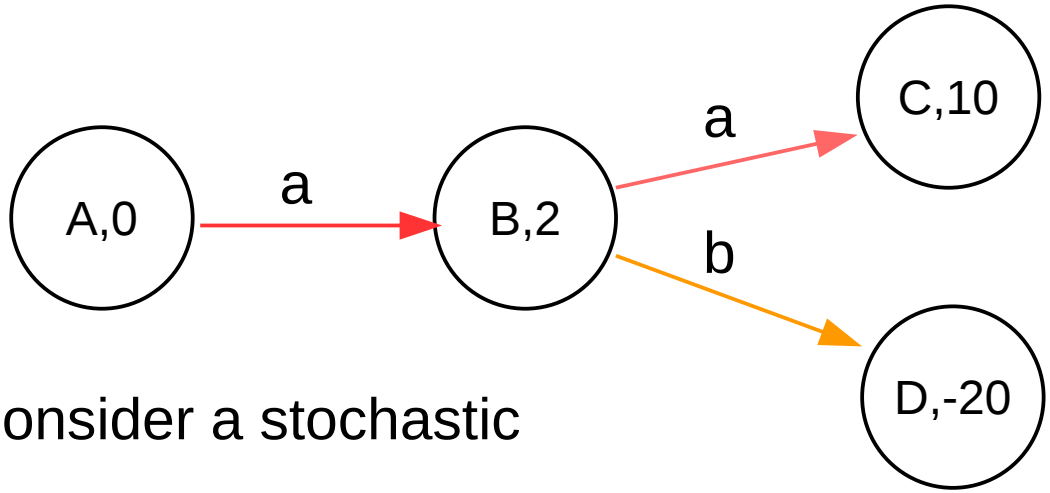
$v_\pi(B) = 10$

$\pi(B) = a$

$v_\pi(A) = 2 + \gamma v_\pi(B) = 7$

Understanding the value

UNIVERSITY OF LEEDS



Now let's consider a stochastic policy:

$\gamma = 0.5$

$\pi(a|B) = 0.8$

$\pi(b|B) = 0.2$

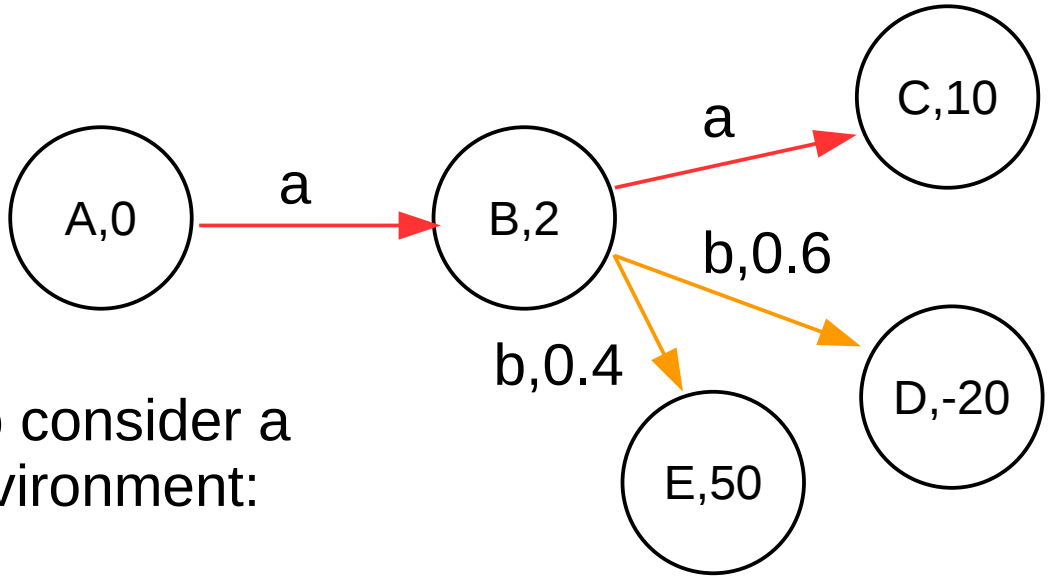
$v_\pi(B) = 0.8 \cdot 10 + 0.2 \cdot (-20) = 4$

$v_\pi(A) = 2 + \gamma\, v_\pi(B) = 4$

A and B are less valuable under this policy, because the agent may go to the bad state.

Understanding the value

UNIVERSITY OF LEEDS



Now let's also consider a stochastic environment:

$\gamma = 0.5$

$\pi(a|B) = 0.8$

$\pi(b|B) = 0.2$

$T(B,b,D) = 0.6$

$T(B,b,E) = 0.4$

$$v_\pi(B) = 0.8 \cdot 10 + 0.2 \cdot (0.6 \cdot (-20) + 0.4 \cdot 50) = 9.6$$

$$v_\pi(A) = 2 + \gamma v_\pi(B) = 6.8$$

Taking b in B may end up in a good state, and the values reflect that.

Understanding the value

UNIVERSITY OF LEEDS

$$v_\pi(B) = 0.8 \cdot 10 + 0.2 \cdot \boxed{(0.6 \cdot (-20) + 0.4 \cdot 50)} = 9.6$$

Probability of taking an action

consequences of action

Breaking down action consequences:

$$(0.6 \cdot (-20 + 0.5 \cdot 0) + 0.4 \cdot (50 + 0.5 \cdot 0))$$

$$(p(D|B,b) \cdot (r(B,b,D) + \gamma v_\pi(D)) + p(E|B,b) \cdot r(B,b,E) + \gamma v_\pi(D))$$

Having chosen an action (b, in this case), the value is the probability of ending up in each possible state, times the value of that state.

Understanding the value

UNIVERSITY OF LEEDS

Probability of taking an action

consequences of action

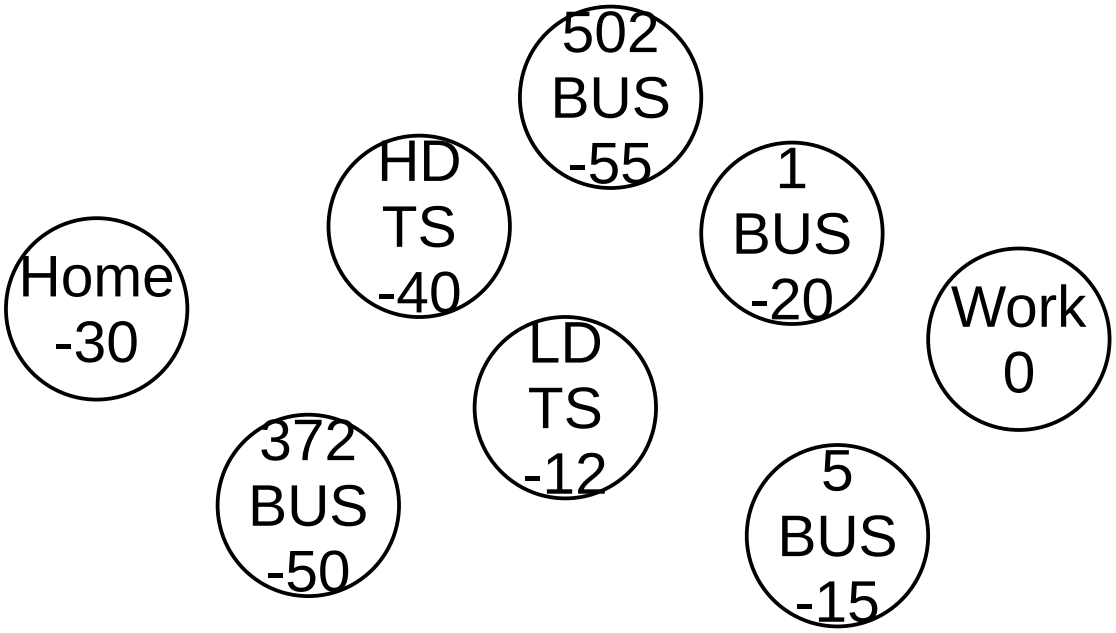
$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)\left(r(s,a,s') + \gamma v_\pi(s')\right)$$

probability of each possible state

value of each possible state

Example: going to work

UNIVERSITY OF LEEDS

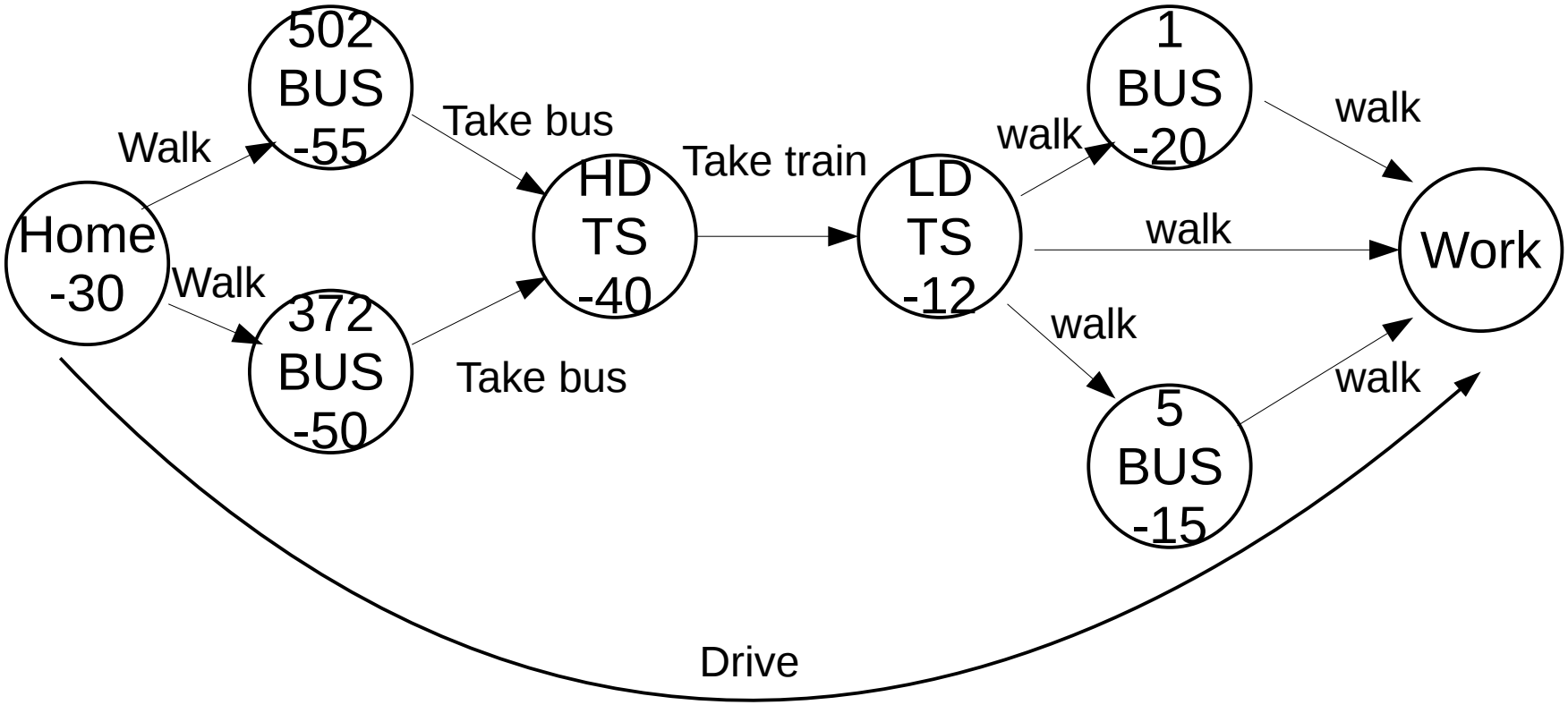


Actions: walk, take bus, take train, drive

How do I retrieve a policy if I only have the values?

Example: going to work

UNIVERSITY OF LEEDS



Action value

UNIVERSITY OF LEEDS

We can also specify the value of a particular action in a given state:

$$q_\pi(s,a)=\sum_{s'} p(s'|s,a)\left(r(s,a,s')+\gamma v_\pi(s')\right)$$

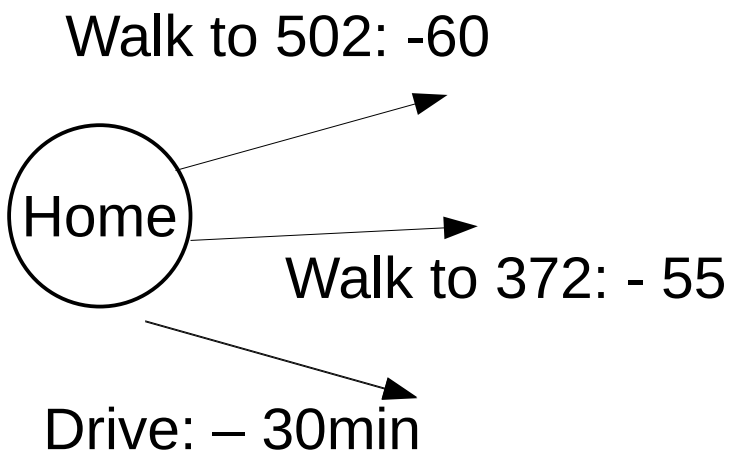
By taking the action out of:

$$v_\pi(s)=\sum_a \pi(a|s)\sum_{s'} p(s'|s,a)\left(r(s,a,s')+\gamma v_\pi(s')\right)$$

From functions to policies

UNIVERSITY OF LEEDS

If we have $q^*(s,a)$



$$\pi^*(s) = argmax_a[q^*(s,a)]$$

Does not require the transition model T!

Optimal value functions

UNIVERSITY OF LEEDS

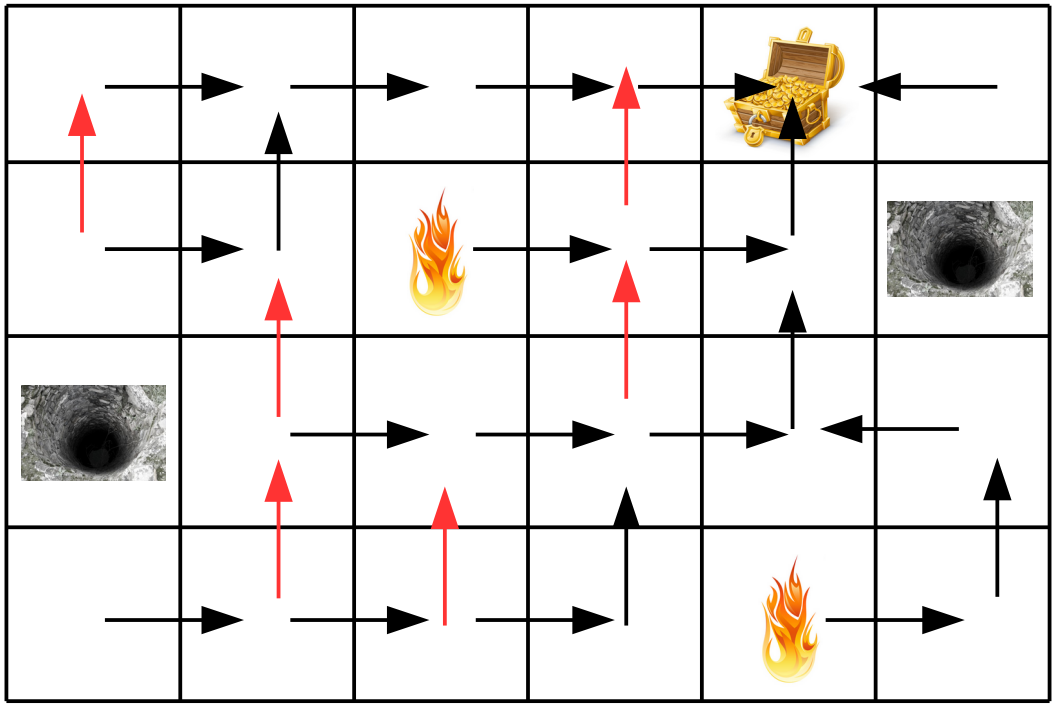
We have now the idea of an *optimal* value function, that is a value function that is the highest it can be on every state, and corresponds to optimal policies.

$$v^*(s) \quad q*(s,a)$$

We now see how to compute the optimal value function

Example: optimal policy

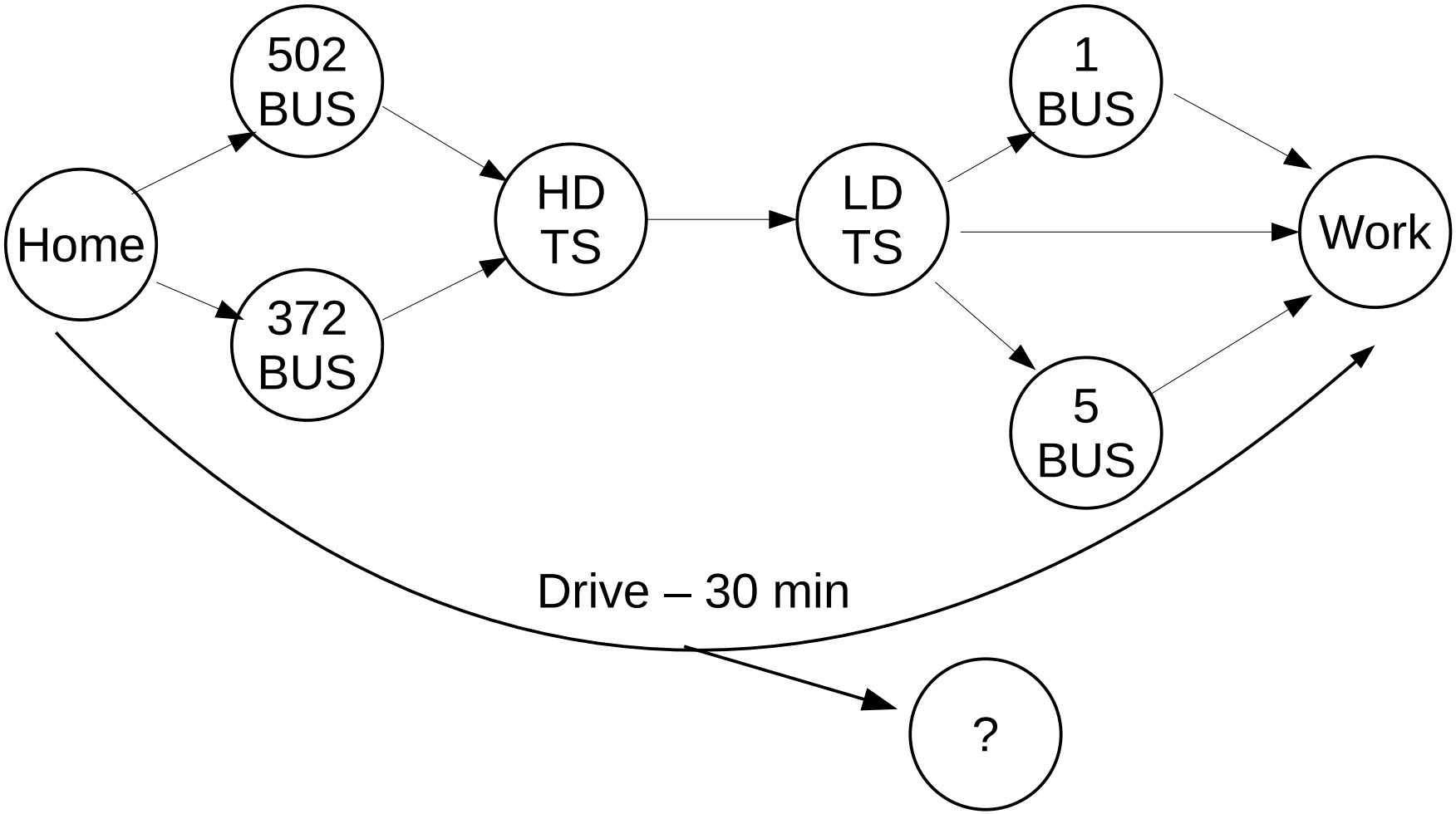
UNIVERSITY OF LEEDS



Why is the optimal policy not unique? You can brake ties between equally valuable actions in many ways (other possible optimal actions in red)

Is this a good model?

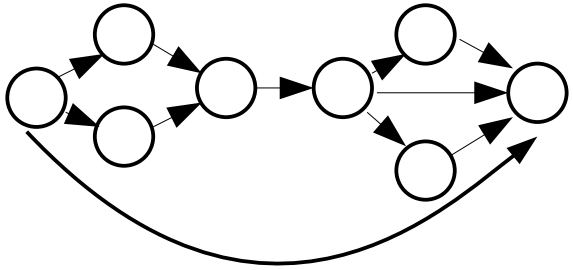
UNIVERSITY OF LEEDS



Dynamic Programming

UNIVERSITY OF LEEDS

If all elements of the MDP are known:



$$MDP\ M = \langle S, A, T, r \rangle$$

We go from the value of any policy π:

$$q_\pi(s,a) = \sum_{s'} p(s'|s,a)(r(s,a,s') + \gamma\, q_\pi(s', \pi(s')))$$

To the value of the *optimal* policy:

$$q^*(s,a) = \sum_{s'} p(s'|s,a)(r(s,a,s') + max_{a'} \gamma\, q^*(s',a'))$$

To the update rule:

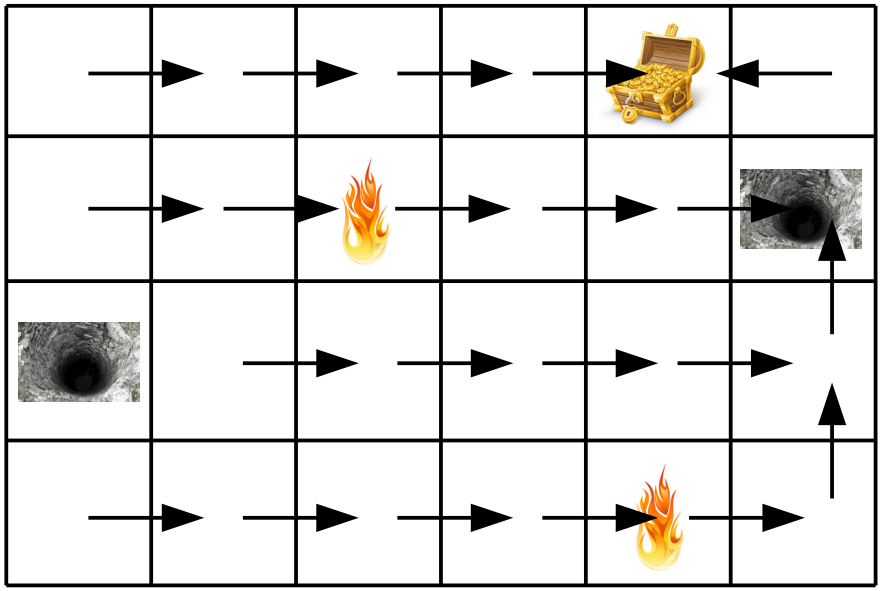
$$q_{k+1}(s,a) = \sum_{s'} p(s'|s,a)(r(s,a,s') + \gamma\, max_{a'} q_k(s',a'))$$

Policy iteration

UNIVERSITY OF LEEDS

Transition and reward function unknown: $MDP\ M = \langle S, A, ?, ? \rangle$

The agent starts with an arbitrary policy (or, equivalently, action-value function), For instance, this policy:

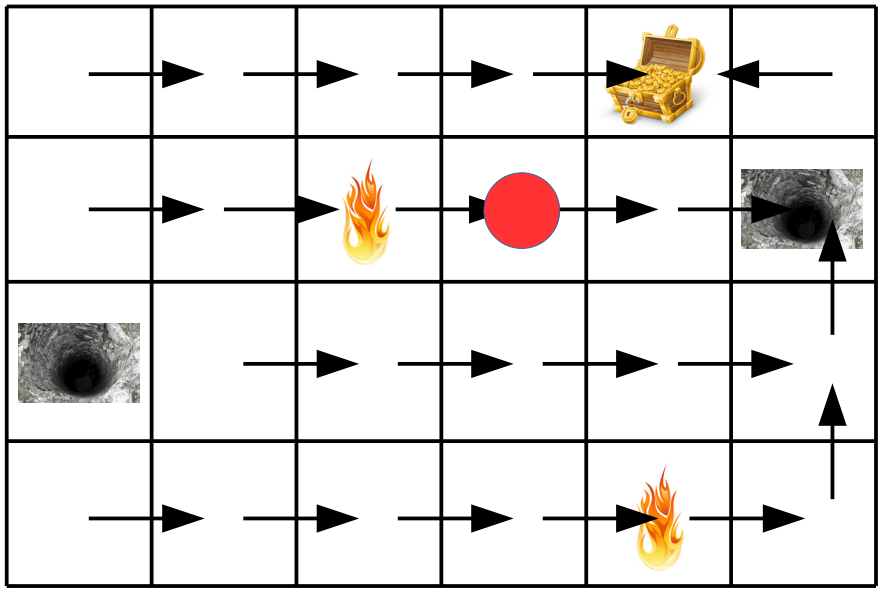


Policy iteration

UNIVERSITY OF LEEDS

Step 1: policy evaluation

Compute the action-value function of this policy



For simplicity, let's focus on one state:

$\gamma = 0.5$

$$q(\langle 4,2 \rangle, \text{right}) = 0 + \gamma(-100) = -50$$

$$q(\langle 4,2 \rangle, \text{up}) = 0 + \gamma(50) = 25$$

$$q(\langle 4,2 \rangle, \text{left}) = -5 + \gamma^3(-100) = -17.5$$

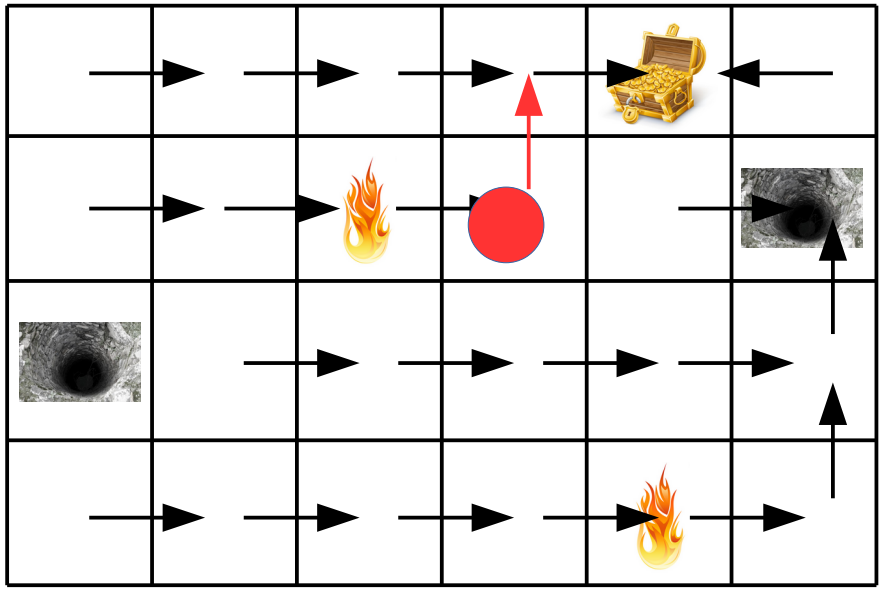
$$q(\langle 4,2 \rangle, \text{down}) = 0 + \gamma^3(-100) = -12.5$$

Policy iteration

UNIVERSITY OF LEEDS

Step 2: policy improvement

Having the value for the current policy, find a state in which a different action would be better than the current one.



$$q(\langle 4,2\rangle, \text{right}) = 0 + \gamma(-100) = -50$$

$$q(\langle 4,2\rangle, \text{up}) = 0 + \gamma(50) = 25$$

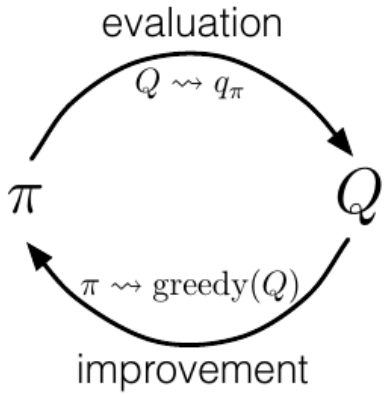
$$q(\langle 4,2\rangle, \text{left}) = -5 + \gamma^3(-100) = -17.5$$

$$q(\langle 4,2\rangle, \text{down}) = 0 + \gamma^3(-100) = -12.5$$

The current action is RIGHT, a better action is UP

Generalized Policy iteration

UNIVERSITY OF LEEDS



No need to fully evaluate a policy before making an improvement
→ *Generalized* PI

On MDPs this is a *hill-climbing* procedure, every solution is strictly better than previous ones.

How do we evaluate a policy without a model?

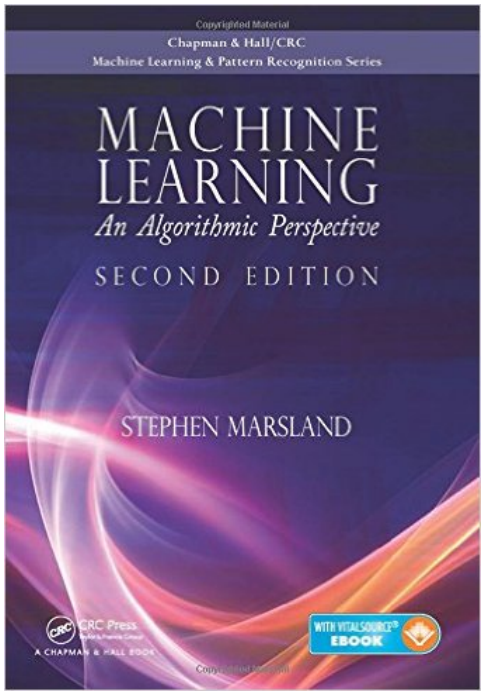
Conclusion

Learning outcomes

UNIVERSITY OF LEEDS

- Define and compute the value and action value function for a given policy
- Perform a step of generalized policy iteration

UNIVERSITY OF LEEDS



Chapter 11