L9: Markov decision processes, value functions and Bellman equations

Lennart Svensson

Department of Electrical Engineering
Chalmers University of Technology, Sweden

Retrieval

- **Motivation:**
  1. retrieval is one of the most efficient tools to strengthen your memory of something
     $\rightsquigarrow$ but it should be slightly "painful"
  2. studies shows that it gives more learning than, e.g., concept maps.

  [J.D. Karpicke and J.R. Blunt: Retrieval Practice Produces More Learning than Elaborative Studying with Concept Mapping. Science, 311, 2011.]
  https://www.youtube.com/watch?v=69VPjsgm-E0

- **Your task:**

(Retrieval) Summarize the content of the videos to yourself, in silence. (2 min)

(Discussion) Explain what you have learned/remember to each other within your groups. (roughly 10 min)

Discussion tasks

- What characterizes reinforcement learning (RL) problems?
- Give two examples of RL problems.
- For at least one of your two problems:
    1. What are possible states, rewards and actions?
    2. Why do we prefer to maximize the value function (the expected return) instead of the reward?
       (Recall that the return is $G_t = R_{t+1} + \gamma R_{t+2} + \dots$)

Discussion tasks (2)

- What characterizes reinforcement learning (RL) problems?
  - instead of supervision in terms of labelled data, we recieve (real valued) rewards,
  - feedback is delayed; it may take a long time before we receive a reward,
  - time sequences, where current decisions affect future states and rewards.

- Try to give two examples of RL problems.

  Four of many possible examples:

  1. Playing chess.
  2. Control a self-driving car.
  3. Domestic robots for household chores.
  4. Develop algorithms to optimize neural networks.

Discussion tasks (3)

Let's discuss the task of playing chess.

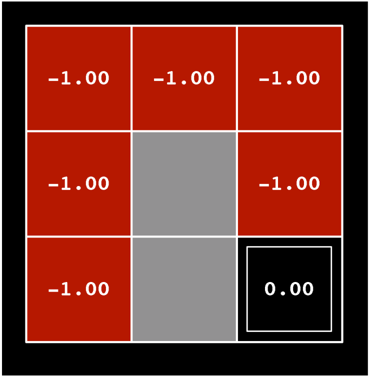
1. What are possible states, rewards and actions?
   - **States**: the position of all pieces on the board.
   - **Rewards:** $+1$ for winning the game, 0 if it's a draw, -1 for losing. During the game rewards are zero, i.e., $R_{t+1} = 0$.
   - **Actions:** how we move one of the pieces.
2. Why do we prefer to maximize the value function (the expected return) instead of the reward?
   (Recall that the return is $G_t = R_{t+1} + R_{t+2} + \dots$)
   - During the game rewards are zero, and it does not make sense to maximize the immediate rewards. We want to win!
   - In chess, we would use $\gamma = 1$ in order not to favor short games.
   - The return is then the final reward for the game ($+1$, 0 or -1), and the value becomes

$$v_\pi(s) = \mathbb{E}\left[G_t \middle| S_t = s\right] = \Pr\left[\text{win} \middle| S_t = s, \pi\right] - \Pr\left[\text{lose} \middle| S_t = s, \pi\right]$$

Credits for the slides

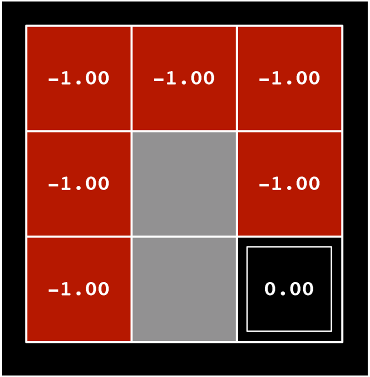
The PIs that follow were originally written by Sébastien Gros, for our joint RL course.

Deterministic environment, moves are (North, South, East, West), move into a wall is blocked, but pay -1!

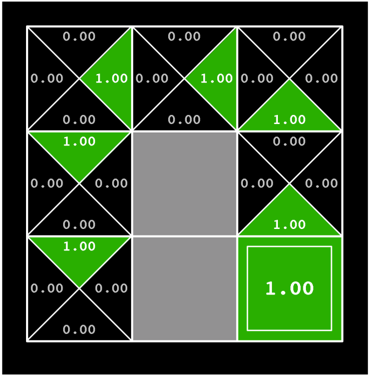


Reward function
$$R(s,a) = \mathbb{E}_\pi\left[R_{t+1}\middle|S_t = s, A_t = a\right]$$

Deterministic environment, moves are (North, South, East, West), move into a wall is blocked, but pay -1!

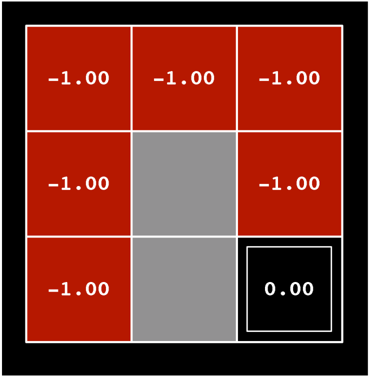


Reward function
$R(s,a) = \mathbb{E}_\pi \left[R_{t+1} \middle| S_t = s, A_t = a \right]$

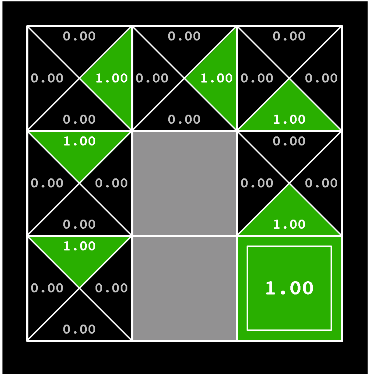


Optimal policy $\pi_\star(a \,|\, s)$

Deterministic environment, moves are (North, South, East, West), move into a wall is blocked, but pay -1!



Reward function
$R(s,a) = \mathbb{E}_\pi \left[R_{t+1} \middle| S_t = s, A_t = a \right]$

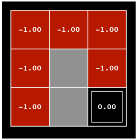


Optimal policy $\pi_\star(a \,|\, s)$

What is the value function $v_\star(s)$??

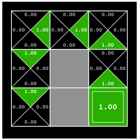
PI: The fabulous Grid World (1)

Deterministic environment, moves are (North, South, East, West), move into a wall is blocked, but pay -1!



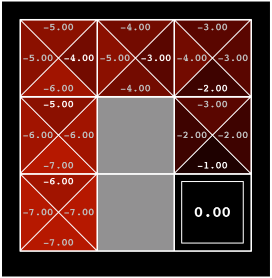
Reward function $R(s, a)$

What is $v_\star(s)$??

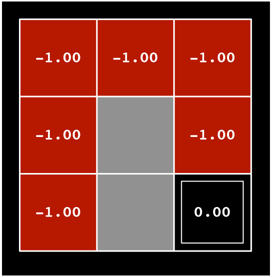


Optimal policy $\pi_\star(a \mid s)$

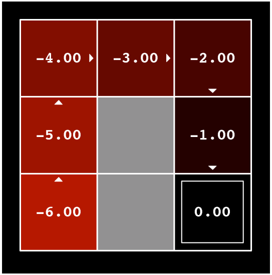
Orange



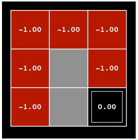
Yellow



Green

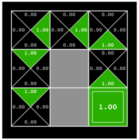


Deterministic environment, moves are (North, South, East, West), move into a wall is blocked, but pay -1!



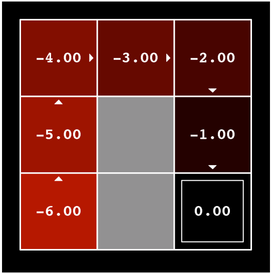
Reward function $R(s, a)$

What is $v_\star(s)$??



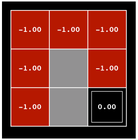
Optimal policy $\pi_\star(a \mid s)$

Green



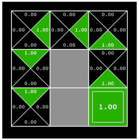
PI: The fabulous Grid World (1)

Deterministic environment, moves are (North, South, East, West), move into a wall is blocked, but pay -1!



Reward function $R(s, a)$

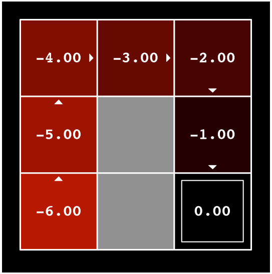
What is $v_\star(s)$??



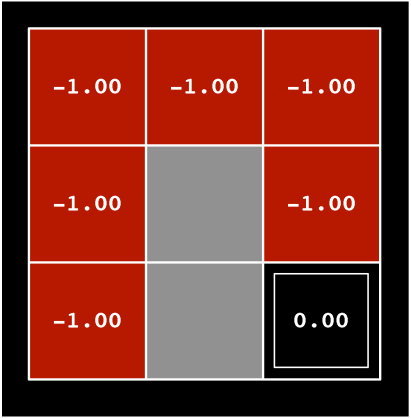
Optimal policy $\pi_\star(a \mid s)$

- Policy: get to the "end box" (2,0) asap
- "Value" of boxes is expected reward to get to "end box". Deterministic problem.
- Value function here is simply the "distance" to the end box!
- Observe: *rewards* (-1 everywhere) v.s. *values* (distance to end)!

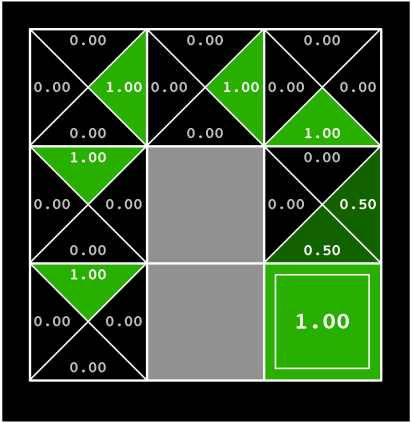
Green



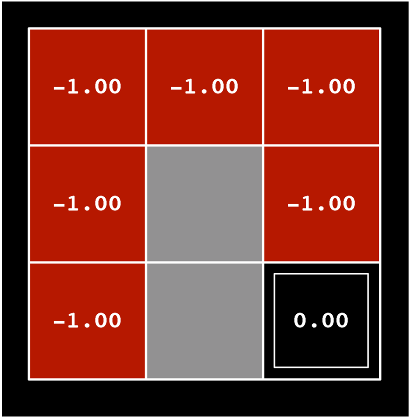
PI: The fabulous Grid World (2)



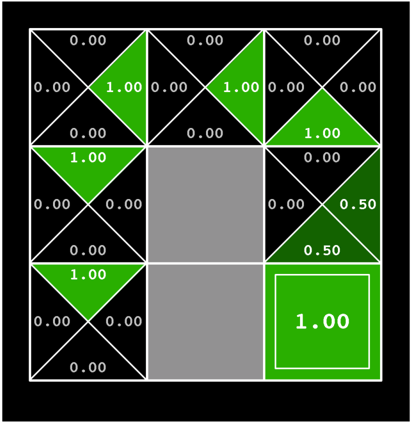
Reward function $R(s, a)$



Policy $\pi(a \mid s)$



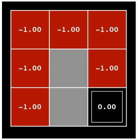
Reward function $R(s, a)$



Policy $\pi(a \mid s)$

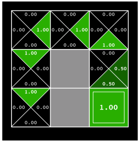
What is the value function $v_\pi(s)$??

PI: The fabulous Grid World (2)



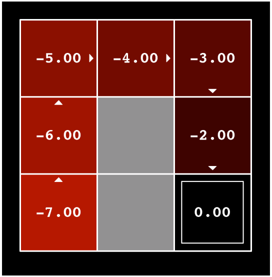
Reward function $R(s, a)$

What is $v_\pi(s)$??

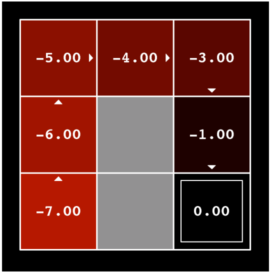


Policy $\pi(a \mid s)$

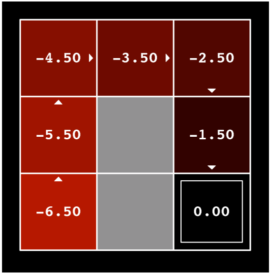
Orange

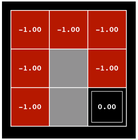


Yellow



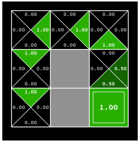
Green





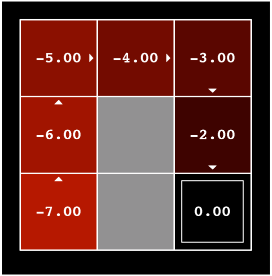
Reward function $R(s, a)$

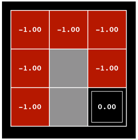
What is $v_\pi(s)$??



Policy $\pi(a \mid s)$

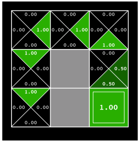
Orange





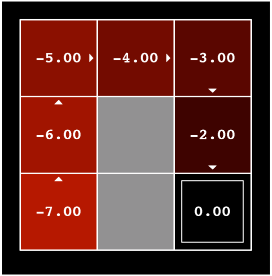
Reward function $R(s, a)$

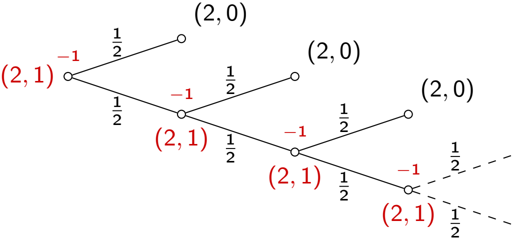
What is $v_\pi(s)$??

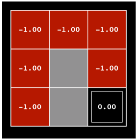


Policy $\pi(a \mid s)$

Orange

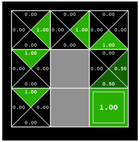






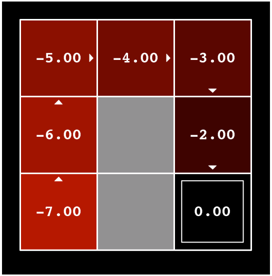
Reward function $R(s, a)$

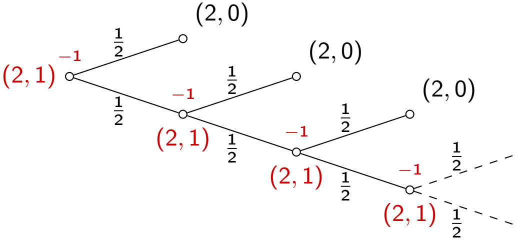
What is $v_\pi(s)$??



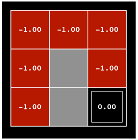
Policy $\pi(a \mid s)$

Orange



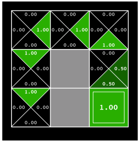


$$v_\pi(\,(2,1)\,) = -1 + \frac{1}{2}\left(-1 + \frac{1}{2}\left(-1 + \ldots\right.\right.$$



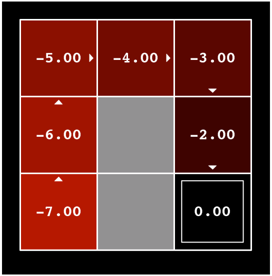
Reward function $R(s, a)$

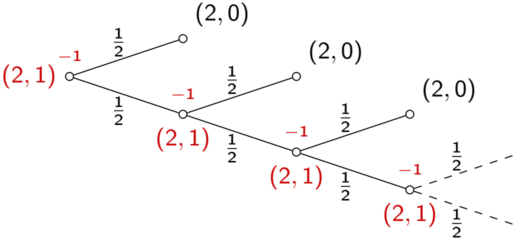
What is $v_\pi(s)$??



Policy $\pi(a \mid s)$

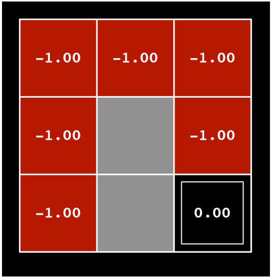
Orange



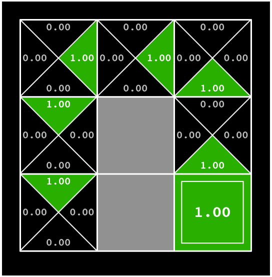


$$v_\pi(\,(2,1)\,) = -1 + \frac{1}{2}\left(-1 + \frac{1}{2}\left(-1 + \ldots \right.\right. = \underbrace{-1 - \frac{1}{2} - \frac{1}{4}\ldots}_{-2}$$

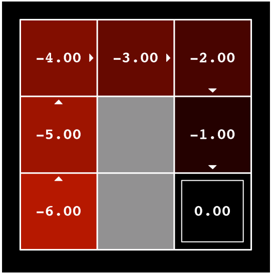
PI: The fabulous Grid World (3)



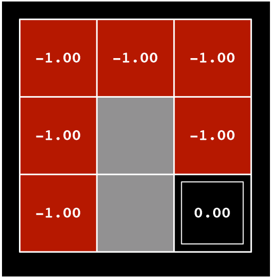
Reward function $R(s, a)$



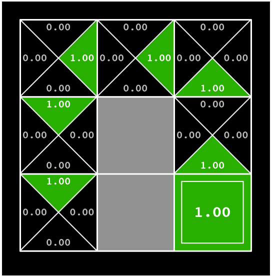
Policy $\pi_\star(a \mid s)$



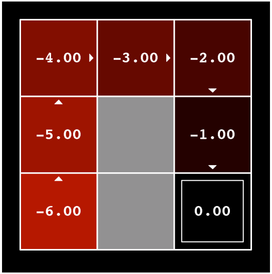
Value function $v_\star(s)$



Reward function $R(s, a)$



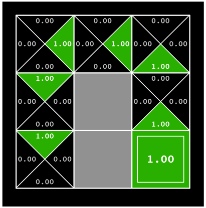
Policy $\pi_\star(a \mid s)$



Value function $v_\star(s)$

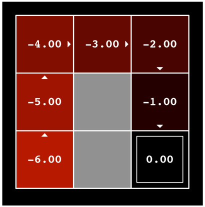
What is the action-value function $q_\star(s, a)$??

PI: The fabulous Grid World (3)



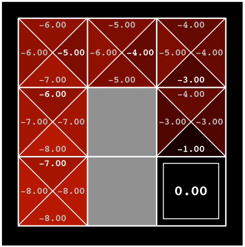
Policy $\pi_\star(a \mid s)$

What is $q_\star(s, a)$??

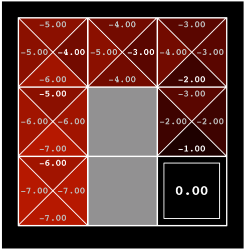


Value function $v_\star(s)$

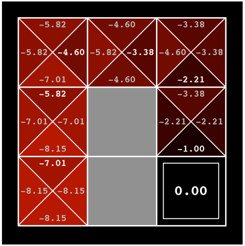
Orange

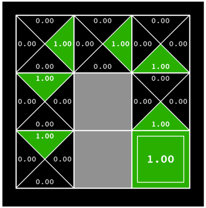


Yellow



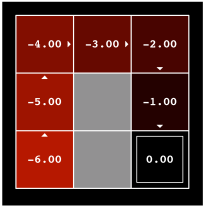
Green





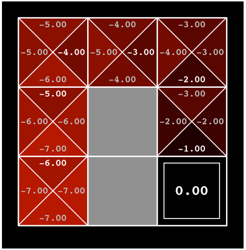
Policy $\pi_\star(a \mid s)$

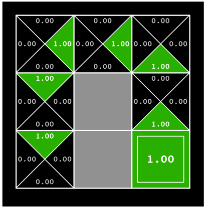
What is $q_\star(s, a)$??



Value function $v_\star(s)$

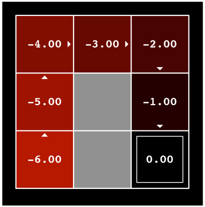
Yellow





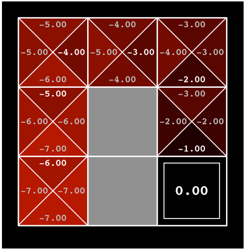
Policy $\pi_\star(a \mid s)$

What is $q_\star(s, a)$??



Value function $v_\star(s)$

Yellow



Optimal action-value function coincides with value functions on optimal actions!!

I.e. $q_\star(s, a_\star) = v_\star(s)$

Bridge or tunnel?

a) Similarly to the previous PIs, we get

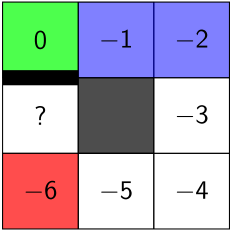


Figure: An illustration of a the value function $v_{\pi_1}(s)$ for the states that we pass on our way to Lindholmen.

- The reward is -1, everywhere except for the terminal state, and the value is simply minus the distance to the end state.

Bridge or tunnel?

b) In this case, the policy is deterministic, and at $s = (2, 0)$ the action is 'North':

$$\pi_1(a|s) = \begin{cases} 1 & \text{if } a = \text{'North'} \\ 0 & \text{otherwise.} \end{cases}$$

Note that deterministic policies are sometimes instead written $\pi(s)$, e.g., $\pi_1((2, 0)) =$'North'.

b) The transition model is also deterministic, and by moving 'North' from $s = (2, 0)$ we end up at $s' = (2, 1)$:

$$\mathcal{P}^{\text{'North'}}_{(2,0)s'} = \Pr\left[S_{t+1} = s' \middle| S_t = (2, 0), A_t = \text{'North'}\right] = \begin{cases} 1 & \text{if } s' = (2, 1) \\ 0 & \text{otherwise.} \end{cases}$$

c) When $s = (2,0)$, $\pi_1(a \big| s)$ is only nonzero for $a =$'North'. Also, if $s = (2,0)$ and $a =$'North', $\mathcal{P}^a_{ss'}$ is only nonzero when $s' = (2,1)$.

- It follows that when $s = (2,0)$,

$$\begin{aligned} v_{\pi_1}(s) &= \sum_a \pi_1(a \big| s) \left(\mathcal{R}^a_s + \gamma \sum_{s'} \mathcal{P}^a_{ss'} v_{\pi_1}(s') \right) \\ &= \mathcal{R}^{\text{'North'}}_{(2,0)} + \gamma v_{\pi_1}((2,1)) \\ &= -1 + v_{\pi_1}((2,1)) \end{aligned}$$

- We already know that $v_{\pi_1}((2,0)) = -4$ and $v_{\pi_1}((2,1)) = -3$, which means that the Bellman equation is satisfied.
- **Interpretation?** Roughly speaking, the value at a state is the immediate reward plus the value at the state where we end up.

d) To find the value function, it is useful to start the calculations from the terminal state. The value at $s = (1, 2)$ is still -1. Whereas the value at $s = (2, 2)$ is

$$\begin{aligned} v_{\pi_1}((2,2)) &= 0.5(-1 + v_{\pi_1}(1,2)) + 0.5(-1 + v_{\pi_1}(0,2)) \\ &= 0.5(-1-1) + 0.5(-1+0) = -1.5. \end{aligned}$$

0	-1	-1.5
?		-2.5
-5.5	-4.5	-3.5

Figure: An illustration of a the value function $v_{\pi_1}(s)$, considering that traffic sometimes improves through the tunnel.

Bridge or tunnel?

d) The expression for transition model, when $s = (2,2)$ and $a =$'West' is

$$\mathcal{P}^a_{ss'} = \begin{cases} 0.5 & \text{if } s' = (0,2) \\ 0.5 & \text{if } s' = (1,2) \\ 0 & \text{otherwise.} \end{cases}$$

- We note that for $s = (2,2)$:

$$\pi_1(a \big| (2,2)) = \begin{cases} 1 & \text{if } a = \text{'West'} \\ 0 & \text{otherwise.} \end{cases}$$

- The Bellman expectation equation when $s = (2,2)$ is

$$\begin{aligned} v_{\pi_1}(s) &= \sum_a \pi_1(a \big| s) \left(\mathcal{R}^a_s + \gamma \sum_{s'} \mathcal{P}^a_{ss'} v_{\pi_1}(s') \right) \\ &= \mathcal{R}^{\text{'West'}}_{(2,2)} + 0.5 v_{\pi_1}((1,2)) + 0.5 v_{\pi_1}((0,2)) \\ &= -1 + 0.5 v_{\pi_1}((1,2)) + 0.5 v_{\pi_1}((0,2)). \end{aligned}$$

- Not only does it match the values we computed, but it exactly matches **how we computed** $v_{\pi_1}((2,2))$.

e) The expected reward for $s = (0,1)$, $a =$'North' is

$$\begin{aligned}\mathcal{R}_{(0,1)}^{\text{'North'}} &= \mathbb{E}\left[R_{t+1} \middle| S_t = (0,1), A_t = \text{'North'}\right] \\ &= 0.8 \times (-1) + 0.2 \times (-10) \\ &= -2.8,\end{aligned}$$

due to the fact that the bridge opens with probability 0.2.

- We still have a deterministic transition function.

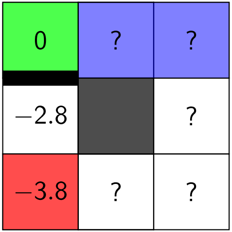


Figure: An illustration of a the value function $v_{\pi_2}(s)$, for $s = (0,2)$, $s = (0,1)$ and $s = (0,0)$.

Bridge or tunnel?

f) We already know that

$$v_{\pi_1}(0,0) = -5.5 \qquad \text{West towards tunnel}$$
$$v_{\pi_2}(0,0) = -3.8 \qquad \text{North across bridge}$$

which means that it is better on average to cross the bridge.

f) The action-value function

$$q_{\pi_1}(s,a) = \mathbb{E}\left[G_t \middle| S_t = s, A_t = a\right]$$

is the expected return (the value) starting from state s, taking action a and then following policy π_1.

f) Comparing this with the definitions of $v_{\pi_1}(0,0)$ and $v_{\pi_2}(0,0)$, we realise that

$$q_{\pi_1}((0,0), \text{'West'}) = v_{\pi_1}(0,0) = -5.5$$
$$q_{\pi_1}((0,0), \text{'North'}) = v_{\pi_2}(0,0) = -3.8.$$

Bridge or tunnel?

f) We already know that

$$\begin{aligned} v_{\pi_1}(0,0) &= -5.5 && \text{West towards tunnel} \\ v_{\pi_2}(0,0) &= -3.8 && \text{North across bridge} \end{aligned}$$

which means that it is better on average to cross the bridge.

f) The action-value function

$$q_{\pi_1}(s,a) = \mathbb{E}\left[G_t \middle| S_t = s, A_t = a\right]$$

is the expected return (the value) starting from state s, taking action a and then following policy π_1.

f) Comparing this with the definitions of $v_{\pi_1}(0,0)$ and $v_{\pi_2}(0,0)$, we realise that

$$\begin{aligned} q_{\pi_1}((0,0), \text{'West'}) &= v_{\pi_1}(0,0) = -5.5 \\ q_{\pi_1}((0,0), \text{'North'}) &= v_{\pi_2}(0,0) = -3.8. \end{aligned}$$

Bridge or tunnel?

f) The greedy strategy for making a decision at $s = (0,0)$ is to solve

$$\arg\max_{a \in \mathcal{A}} q_{\pi_1}((0,0), a)$$

which tells us to cross the bridge since that gives a larger action-value (-3.8).

f) The action-value function

$$q_{\pi_1}(s,a) = \mathbb{E}\left[G_t \middle| S_t = s, A_t = a\right]$$

is the expected return (the value) starting from state s, taking action a and then following policy π_1.

f) Comparing this to the definitions of $v_{\pi_1}(0,0)$ and $v_{\pi_2}(0,0)$, we realise that

$$\begin{aligned} q_{\pi_1}((0,0), \text{'West'}) &= v_{\pi_1}(0,0) = -5.5 \\ q_{\pi_1}((0,0), \text{'North'}) &= v_{\pi_2}(0,0) = -3.8. \end{aligned}$$

Bridge or tunnel?

f) We can also use the Bellman expectation equation to express q_π in terms of v_π:

$$q_{\pi_1}(s,a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_{\pi_1}(s'), \tag{1}$$

to reach the same conclusion.

- The fact that the transition model is deterministic for $s = (0,0)$ simplifies the expression considerably.
- For instance, for $s = (0,0)$ and $a =$'West, we get

$$\begin{aligned} q_{\pi_1}(s,a) &= \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_{\pi_1}(s') \\ &= -1 + v_{\pi_1}((1,0)) \\ &= -1 - 4.5 = -5.5. \end{aligned}$$

- Similarly, $q_{\pi_1}((0,0), \text{'North'}) = -1 + v_{\pi_1}((0,1)) = -1 - 2.8 = -3.8$.