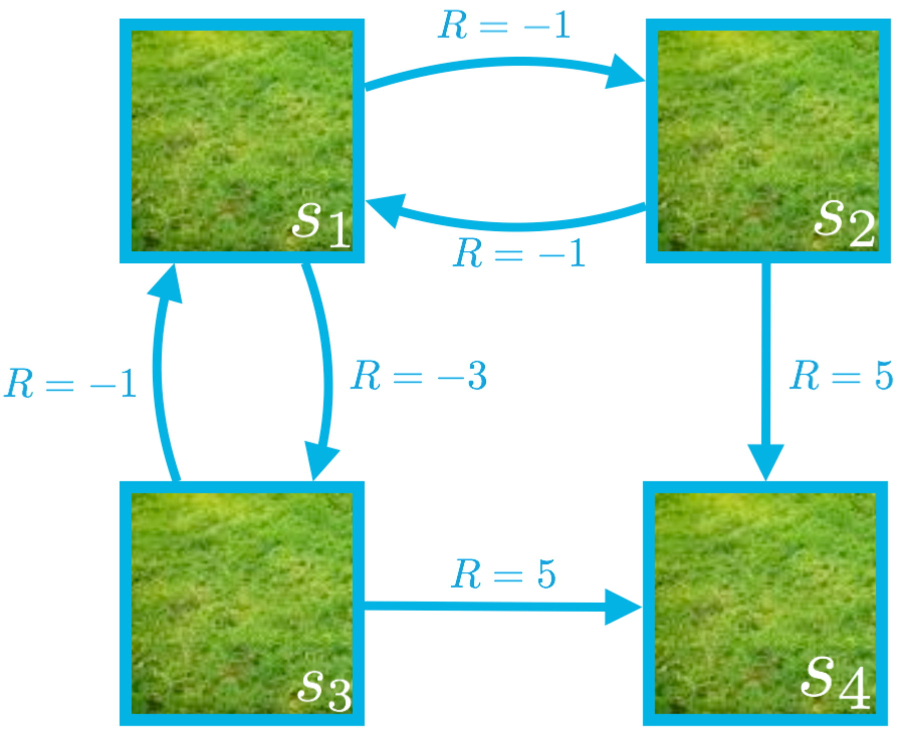
An Iterative Method

In this concept, we will examine some ideas from the last video in more detail.



Notes on the Bellman Expectation Equation

In the previous video, we derived one equation for each environment state. For instance, for state s_1, we saw that:

$$v_\pi(s_1) = \frac{1}{2}(-1 + v_\pi(s_2)) + \frac{1}{2}(-3 + v_\pi(s_3)).$$

We mentioned that this equation follows directly from the Bellman expectation equation for v_π.

> $v_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s] = \sum_{a\in\mathcal{A}(s)}\pi(a|s)\sum_{s'\in\mathcal{S},r\in\mathcal{R}}p(s',r|s,a)(r +$
> (**The Bellman expectation equation for** v_π)

$v_\pi(s_1) = \sum_{a\in\mathcal{A}(s_1)} \pi(a|s_1) \sum_{s'\in\mathcal{S}, r\in\mathcal{R}} p(s', r|s_1, a)(r + \gamma v_\pi(s'))$

Then, it's possible to derive the equation for state s_1 by using the following:

- $\mathcal{A}(s_1) = \{\text{down}, \text{right}\}$ (*When in state s_1, the agent only has two potential actions: down or right.*)
- $\pi(down|s_1) = \pi(\text{right}|s_1) = \frac{1}{2}$ (*We are currently examining the policy where the agent goes down with 50% probability and right with 50% probability when in state s_1.*)
- $p(s_3, -3|s_1, \text{down}) = 1$ (and $p(s', r|s_1, \text{down}) = 0$ if $s' \neq s_3$ or $r \neq -3$) (*If the agent chooses to go down in state s_1, then with 100% probability, the next state is s_3, and the agent receives a reward of -3.*)
- $p(s_2, -1|s_1, \text{right}) = 1$ (and $p(s', r|s_1, \text{right}) = 0$ if $s' \neq s_2$ or $r \neq -1$) (*If the agent chooses to go right in state s_1, then with 100% probability, the next state is s_2, and the agent receives a reward of -1.*)
- $\gamma = 1$ (*We chose to set the discount rate to 1 in this gridworld example.*)

If this is not entirely clear to you, please take the time now to plug in the values to derive the equation from the video. Then, you are encouraged to repeat the same process for the other states.

Notes on Solving the System of Equations

In the video, we mentioned that you can directly solve the system of equations:

$v_\pi(s_1) = \frac{1}{2}(-1 + v_\pi(s_2)) + \frac{1}{2}(-3 + v_\pi(s_3))$

$v_\pi(s_2) = \frac{1}{2}(-1 + v_\pi(s_1)) + \frac{1}{2}(5 + v_\pi(s_4))$

$v_\pi(s_3) = \frac{1}{2}(-1 + v_\pi(s_1)) + \frac{1}{2}(5 + v_\pi(s_4))$

$v_\pi(s_4) = 0$

Since the equations for $v_\pi(s_2)$ and $v_\pi(s_3)$ are identical, we must have that $v_\pi(s_2) = v_\pi(s_3)$.

Thus, the equations for $v_\pi(s_1)$ and $v_\pi(s_2)$ can be changed to:

$v_\pi(s_2) = \frac{1}{2}(-1 + v_\pi(s_1)) + \frac{1}{2}(5 + 0) = 2 + \frac{1}{2}v_\pi(s_1)$

Combining the two latest equations yields

$v_\pi(s_1) = -2 + 2 + \frac{1}{2}v_\pi(s_1) = \frac{1}{2}v_\pi(s_1)$,

which implies $v_\pi(s_1) = 0$. Furthermore, $v_\pi(s_3) = v_\pi(s_2) = 2 + \frac{1}{2}v_\pi(s_1) = 2 + 0 = 2$.

Thus, the state-value function is given by:

$v_\pi(s_1) = 0$

$v_\pi(s_2) = 2$

$v_\pi(s_3) = 2$

$v_\pi(s_4) = 0$

Note. This example serves to illustrate the fact that it is ***possible*** to *directly* solve the system of equations given by the Bellman expectation equation for v_π. However, in practice, and especially for much larger Markov decision processes (MDPs), we will instead use an *iterative* solution approach.

NEXT