# The Bellman Update

$$U_{i+1}(s) \leftarrow R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s' \mid s, a) U_i(s')$$

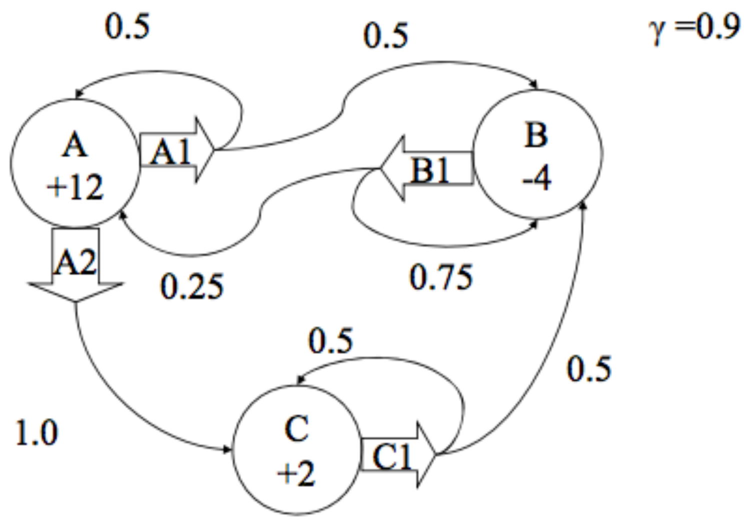
This is the maximum possible expected sum of discounted rewards (utilities) if the agent is at state s and lives for i+1 time steps.

Apply the Bellman update until the utility function converges.

The optimal policy is given by:

$$\pi^*(s) = \operatorname*{argmax}_{a \in A(s)} \sum_{s'} P(s' \mid s, a) U(s')$$

# Example



# Example

i=1

$U_1(A) = R(A)=12$

$U_1(B) = R(B)=-4$

$U_1(C) = R(C)=2$

# Example

| $U_1(A)$ | $U_1(B)$ | $U_1(C)$ |
|---|---|---|
| 12 | -4 | 2 |

i=2

$U_2(A)$ = 12 + (0.9) * max{(0.5)(12)+(0.5)(-4), (1.0)(2)} = 12 + (0.9)*max{4.0,2.0} = 12 + 3.6 = 15.6

$U_2(B)$ = -4 + (0.9) * {(0.25)(12)+(0.75)(-4)} = -4 + (0.9)*0 = -4

$U_2(C)$ = 2 + (0.9) * {(0.5)(2)+(0.5)(-4)} = 2 + (0.9)*(-1) = 2-0.9 = 1.1

# Example

| $U_2(A)$ | $U_2(B)$ | $U_2(C)$ |
|---|---|---|
| 15.6 | -4 | 1.1 |

i=3

$U_3(A)$ = 12 + (0.9) * max{(0.5)(15.6)+(0.5)(-4),(1.0)(1.1)} = 12 + (0.9) * max{5.8,1.1} = 12 + (0.9)(5.8) = 17.22

$U_3(B)$ = -4 + (0.9) * {(0.25)(15.6)+(0.75)(-4)} = -4 + (0.9)*(3.9-3) = -4 + (0.9)(0.9) = -3.19

$U_3(C)$ = 2 + (0.9) * {(0.5)(1.1)+(0.5)(-4)} = 2 + (0.9)*{0.55-2.0} = 2 + (0.9)(-1.45) = 0.695

## Value-Iteration Termination

When do you stop?

In an iteration over all the states, keep track of the maximum change in utility of any state (call this $\delta$)

When $\delta$ is less than some pre-defined threshold, stop

This will give us an approximation to the true utilities, we can act greedily based on the approximated state utilities