

Value Iteration

Declare:

U', U - current utilities for states, initially \emptyset

δ - max change of any utility of any state in an iteration

ϵ - max allowable error of utility in any state. (controls convergence of algorithm)

Requires MDP with:

Actions, $a \in A(s)$: actions in state s

States S

Rewards $R(s)$: Rewards in state s

discount γ

Transition Model : $P(s' | a, s)$, aka $T_a(s, a, s')$

Loop

$U = U'$

$\delta = \emptyset$

for each state in States

$$U'(s) = R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s' | a, s) U(s')$$

if $\text{abs}(U'(s) - U(s)) > \delta$



$$\delta = \text{abs}(U'(s) - U(s))$$

Until $\delta < \epsilon(1 - \gamma) / \gamma$

$$U_{1,2}(2,3) = R(2,3) + \gamma \max \left[\begin{array}{l} \rightarrow (-1 \times .8 + .72 \times .1 + .1 \times 0), \quad \uparrow (.72 \times .8 + -1 \times .1 + 0 \times .1) \\ \downarrow (0 \times .8 + .1 \times -1 + .1 \times 0), \quad \leftarrow (.8 \times 0 + .1 \times 0 + .1 \times 0) \end{array} \right]$$

$$\gamma \max \left[\begin{array}{l} \rightarrow -0.728, \quad \uparrow .4284 \\ \downarrow -0.1, \quad \leftarrow 0 \end{array} \right]$$

$$= 0 + .9 \times .4284 = .4284$$

		$\rightarrow .72$	+1
		$\uparrow .4284$	-1
			

(4)

Fill in values for all states for $U_1(s) \forall s$

~~Restart~~



Util, ties are closer for states closer to the exit.
because fewer steps are required.

Is (1,2)

Same as (2,3). It's

Same # of steps?

Shouldn't be, there's
no fire, it's danger.

	?	$\rightarrow .72$	+1	1
		$\uparrow .4284$	-1	2 (x)
				3
1	2	3	4	

(4)

How long does value iteration run?

Until

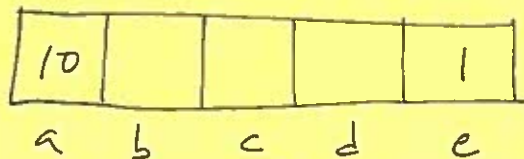
convergence - when ut.lities for ~~any~~ ^{all} states change less than δ .

Value iteration gives us the ut.lities, can also store the actions, or go back through the grid and move toward highest ut.lity

Exercises

12

Given:



is reward

Actions: East, West, and Exit (in states a, e only)

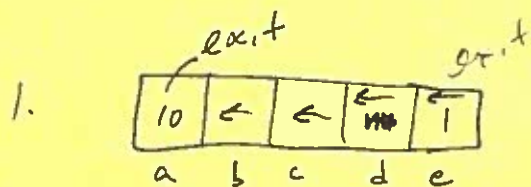
Transitions: deterministic

1. For $\gamma = 1$, what is optimal policy? Action for each state

2. For $\gamma = .1$, what is optimal policy?

3. For what γ are E and W equally good when in state d?

$$V(s) = R(s) + \gamma \max(\leftarrow, \rightarrow)$$

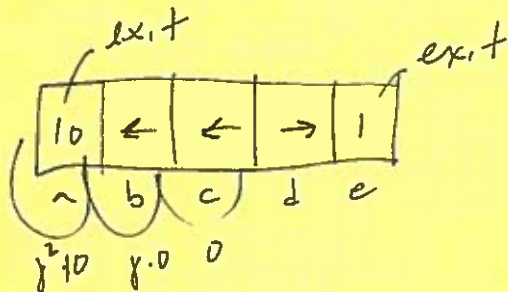


no discounting, go to highest reward

$$\text{from d, } \gamma^3 \times 10 = 10$$

2. Go west for all and get (from d). Go east from d:
 $\gamma^3 \times 10 = .1^3 \times 10 = .01$
 $\gamma \times 1 = .1$ (better)

$$\text{from c } \gamma^2 \times 10 = .01 \times 10 = .1$$



$$.01 \times 10 = .1$$

3. $\gamma^3 \times 10 = \gamma \times 1.0 = \gamma^2 \times 10 = 1$
 $\gamma = \sqrt{\frac{1}{10}}$