

Policy Iteration for Navigation

The little prince lives on a torus, a donut shaped planet divided into nine regions (states)

Here's a map of the world, showing the rewards for each state.

| | | |
|------------------------|----------------------|---------------------------|
| -1 <i>a</i> | -1 <i>b</i> | great food 10 <i>c</i> |
| -1 <i>d</i> | sand - 5 <i>e</i> | monster - 4 <i>f</i> |
| avg food 5 <i>g</i> | -1 <i>h</i> | -1 <i>i</i> |

Let's assume a discount factor γ of .9.

In our general terminology,

$$R = \begin{bmatrix} -1 \\ -1 \\ 10 \\ -1 \\ -5 \\ -4 \\ 5 \\ -1 \\ -1 \end{bmatrix}$$

At each step, the LP has a choice to go N, S, E, or W. He cannot stay stationary.

Since the planet is a torus (donut), going N in state *a* leads go *g*, going W leads to *c*, etc.

So the actions are always *N*, *S*, *E*, *W*. We let *a* stand for one of these.

However, they don't always work.

If you choose *N*, with prob .8 you go *N*, with prob .1 you go *E*, and with prob .1 you go *W*.

If you choose *E*, with prob .8 you go *E*, with prob .1 you go *N*, and with prob .1 you go *S*.

If you choose *S*, with prob .8 you go *S*, with prob .1 you go *E*, and with prob .1 you go *W*.

If you choose *W*, with prob .8 you go *W*, with prob .1 you go *N*, and with prob .1 you go *S*.

1 The first policy

We take π_0 to be the policy that says “go north” in every state. We have

$$I - \gamma P^{\pi_0} = \begin{bmatrix} 1 & -.09 & -.09 & 0 & 0 & 0 & -.72 & 0 & 0 \\ -.09 & 1 & -.09 & 0 & 0 & 0 & 0 & -.72 & 0 \\ -.09 & -.09 & 1 & 0 & 0 & 0 & 0 & 0 & -.72 \\ -.72 & 0 & 0 & 1 & -.09 & -.09 & 0 & 0 & 0 \\ 0 & -.72 & 0 & -.09 & 1 & -.09 & 0 & 0 & 0 \\ 0 & 0 & -.72 & -.09 & -.09 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -.72 & 0 & 0 & 1 & -.09 & -.09 \\ 0 & 0 & 0 & 0 & -.72 & 0 & -.09 & 1 & -.09 \\ 0 & 0 & 0 & 0 & 0 & -.72 & -.09 & -.09 & 1 \end{bmatrix}$$

$$(I - \gamma P^{\pi_0})^{-1} = \begin{bmatrix} 2.118 & 0.829 & 0.829 & 1.345 & 0.783 & 0.783 & 1.672 & 0.821 & 0.821 \\ 0.829 & 2.118 & 0.829 & 0.783 & 1.345 & 0.783 & 0.821 & 1.672 & 0.821 \\ 0.829 & 0.829 & 2.118 & 0.783 & 0.783 & 1.345 & 0.821 & 0.821 & 1.672 \\ 1.672 & 0.821 & 0.821 & 2.118 & 0.829 & 0.829 & 1.345 & 0.783 & 0.783 \\ 0.821 & 1.672 & 0.821 & 0.829 & 2.118 & 0.829 & 0.783 & 1.345 & 0.783 \\ 0.821 & 0.821 & 1.672 & 0.829 & 0.829 & 2.118 & 0.783 & 0.783 & 1.345 \\ 1.345 & 0.783 & 0.783 & 1.672 & 0.821 & 0.821 & 2.118 & 0.829 & 0.829 \\ 0.783 & 1.345 & 0.783 & 0.821 & 1.672 & 0.821 & 0.829 & 2.118 & 0.829 \\ 0.783 & 0.783 & 1.345 & 0.821 & 0.821 & 1.672 & 0.829 & 0.829 & 2.118 \end{bmatrix}$$

$$V^{\pi_0} = \begin{bmatrix} 3.672 \\ -3.686 \\ 11.054 \\ 1.301 \\ -7.229 \\ 3.426 \\ 5.567 \\ -5.572 \\ 1.466 \end{bmatrix}$$

2 The next policy π_1

Then we use V^{π_0} to define π_1 .

| s | north | south | east | west | argmax |
|-----|--------|--------|--------|--------|--------|
| a | 5.190 | 1.778 | -2.262 | 9.530 | west |
| b | -2.985 | -4.311 | 7.563 | 1.658 | east |
| c | 1.171 | 2.739 | 3.427 | -2.460 | east |
| d | 2.557 | 4.073 | -4.859 | 3.665 | south |
| e | -2.476 | -3.985 | 1.815 | 0.115 | east |
| f | 8.250 | 0.580 | 2.293 | -4.531 | north |
| g | 0.630 | 2.527 | -3.960 | 1.670 | south |
| h | -5.080 | -2.246 | 0.081 | 3.362 | west |
| i | 2.740 | 8.843 | 5.902 | -3.010 | south |

The next policy π_1 is given by the rightmost column. This policy gives us a transition matrix P^{π_1} . $I - \gamma P^{\pi_1}$ is shown below:

$$I - \gamma P^{\pi_1} = \begin{bmatrix} 1 & 0 & -.72 & -.09 & 0 & 0 & -.09 & 0 & 0 \\ 0 & 1 & -.72 & 0 & -.09 & 0 & 0 & -.09 & 0 \\ -.72 & 0 & 1 & 0 & 0 & -.09 & 0 & 0 & -.09 \\ 0 & 0 & 0 & 1 & -.09 & -.09 & -.72 & 0 & 0 \\ 0 & -.09 & 0 & 0 & 1 & -.72 & 0 & -.09 & 0 \\ 0 & 0 & -.72 & -.09 & -.09 & 1 & 0 & 0 & 0 \\ -.72 & 0 & 0 & 0 & 0 & 0 & 1 & -.09 & -.09 \\ 0 & -.09 & 0 & 0 & -.09 & 0 & -.72 & 1 & 0 \\ 0 & 0 & -.72 & 0 & 0 & 0 & -.09 & -.09 & 1 \end{bmatrix}$$

The new value vector V^{π_1} is

$$V^{\pi_1} = \begin{bmatrix} 32.692 \\ 31.536 \\ 39.049 \\ 27.944 \\ 20.906 \\ 28.512 \\ 34.022 \\ 28.216 \\ 32.717 \end{bmatrix}$$

Then we make all the “argmax” calculations. The changes to the policy are indicated in UPPER CASE.

| s | north | south | east | west | argmax |
|-----|--------|--------|--------|--------|--------|
| a | 34.276 | 29.414 | 31.425 | 37.436 | west |
| b | 29.747 | 23.899 | 36.151 | 31.066 | east |
| c | 32.596 | 29.232 | 32.277 | 31.352 | NORTH |
| d | 31.095 | 32.159 | 23.396 | 29.481 | south |
| e | 30.874 | 28.218 | 28.785 | 28.330 | NORTH |
| f | 36.124 | 31.059 | 29.532 | 23.901 | north |
| g | 28.449 | 32.247 | 28.636 | 32.237 | south |
| h | 23.399 | 31.903 | 31.418 | 32.462 | west |
| i | 29.033 | 37.463 | 33.974 | 29.329 | south |

3 The next policy π_2

This defines a policy π_2 . We show $I - \gamma P^{\pi_2}$ below:

$$\begin{bmatrix} 0 & 0 & -.72 & -.09 & 0 & 0 & -.09 & 0 & 0 \\ 0 & 0 & -.72 & 0 & -.09 & 0 & 0 & -.09 & 0 \\ -.09 & -.09 & 0 & 0 & 0 & 0 & 0 & 0 & -.72 \\ 0 & 0 & 0 & 0 & -.09 & -.09 & -.72 & 0 & 0 \\ 0 & -.72 & 0 & -.09 & 0 & -.09 & 0 & 0 & 0 \\ 0 & 0 & -.72 & -.09 & -.09 & 0 & 0 & 0 & 0 \\ -.72 & 0 & 0 & 0 & 0 & 0 & 0 & -.09 & -.09 \\ 0 & -.09 & 0 & 0 & -.09 & 0 & -.72 & 0 & 0 \\ 0 & 0 & -.72 & 0 & 0 & 0 & -.09 & -.09 & 0 \end{bmatrix}$$

And this now has an long-term expected discounted total value

$$V^{\pi_2} = \begin{bmatrix} 33.891 \\ 32.918 \\ 40.432 \\ 29.123 \\ 24.012 \\ 29.893 \\ 35.100 \\ 29.395 \\ 33.916 \end{bmatrix}$$

Then we make all the “argmax” calculations.

| s | north | south | east | west | argmax |
|---|--------|--------|--------|--------|--------|
| a | 35.415 | 30.633 | 32.757 | 38.768 | west |
| b | 30.948 | 26.642 | 37.686 | 32.454 | east |
| c | 33.814 | 30.595 | 33.494 | 32.715 | north |
| d | 32.503 | 33.471 | 26.109 | 30.814 | south |
| e | 32.236 | 29.418 | 30.146 | 29.530 | north |
| f | 37.659 | 32.446 | 30.733 | 26.644 | north |
| g | 29.630 | 33.444 | 29.817 | 33.434 | south |
| h | 26.111 | 33.236 | 32.826 | 33.773 | west |
| i | 30.364 | 38.795 | 35.113 | 30.549 | south |

4 π_2 is unimprovable

In all states, the action that leads to the maximum value is the same one as in π_2 . So we declare π_2 to be our **unimprovable** policy. By results to be shown in the homework, π_2 then gives the **optimal** policy, and V^{π_2} as shown above is V^* for this MDP.

This policy π_2 is shown graphically below:

| | | |
|--|------------------------------------|---|
| -1 a $\Leftarrow 37.8$ | -1 b $\Rightarrow 32.5$ | great food + 10 c $\Uparrow 38.8$ |
| -1 d $\Downarrow 30.8$ | sand - 5 e $\Uparrow 29.5$ | monster - 4 f $\Uparrow 26.6$ |
| avg food + 5 g $\Downarrow 33.4$ | -1 h $\Leftarrow 33.8$ | -1 i $\Downarrow 30.5$ |