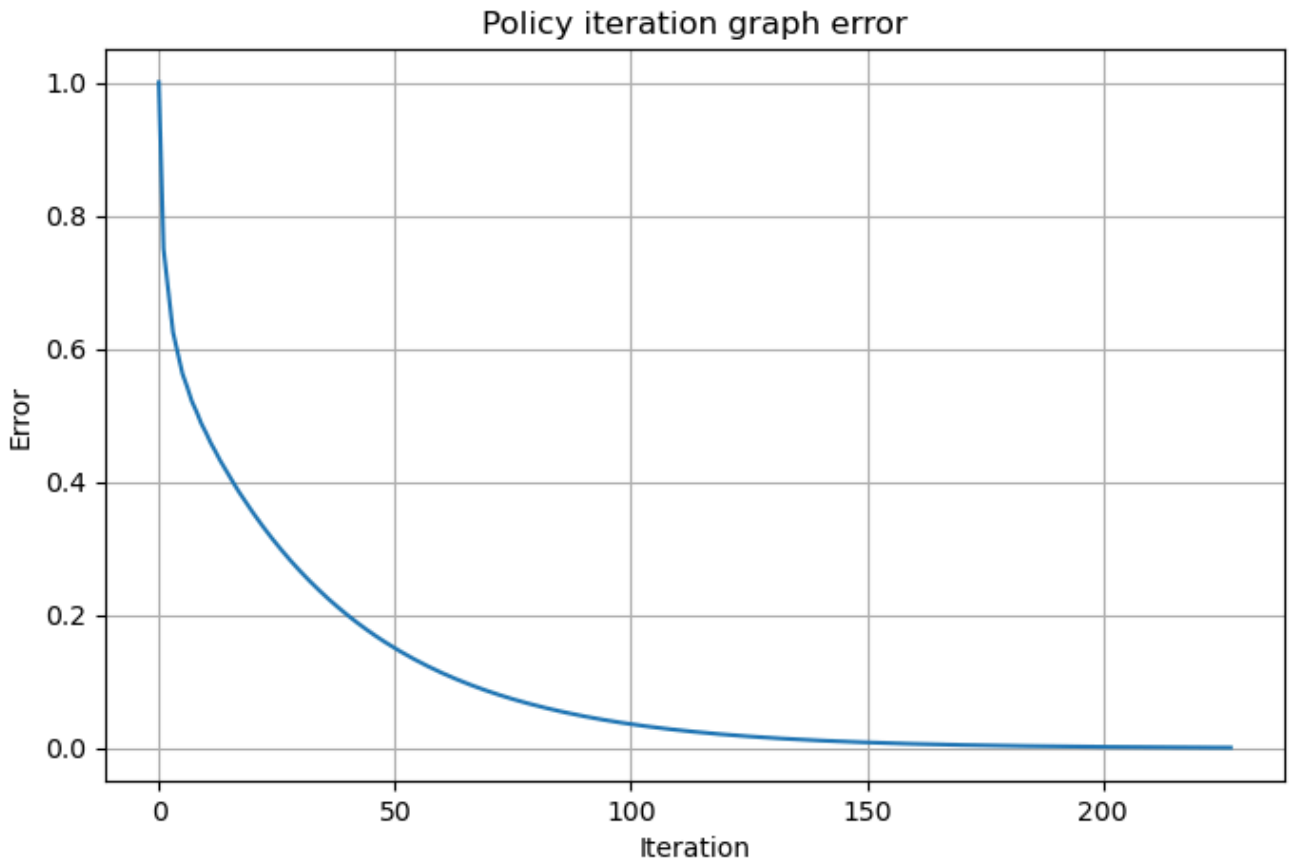
output

Policy iteration:
Starting
Finished after 4 iterations and 0.0005717277526855469 seconds

Done!



Best Actions:
. < < < V
^ ^ ^ ^ V
^ ^ ^ > V
^ ^ > > V
^ > > > .
Score board
[[0. 0. -1. -1.25 -1.3125]
[0. -1. -1.25 -1.3125 -1.25]
[-1. -1.25 -1.3125 -1.25 -1.]
[-1.25 -1.3125 -1.25 -1. 0.]
[-1.3125 -1.25 -1. 0. 0.]]

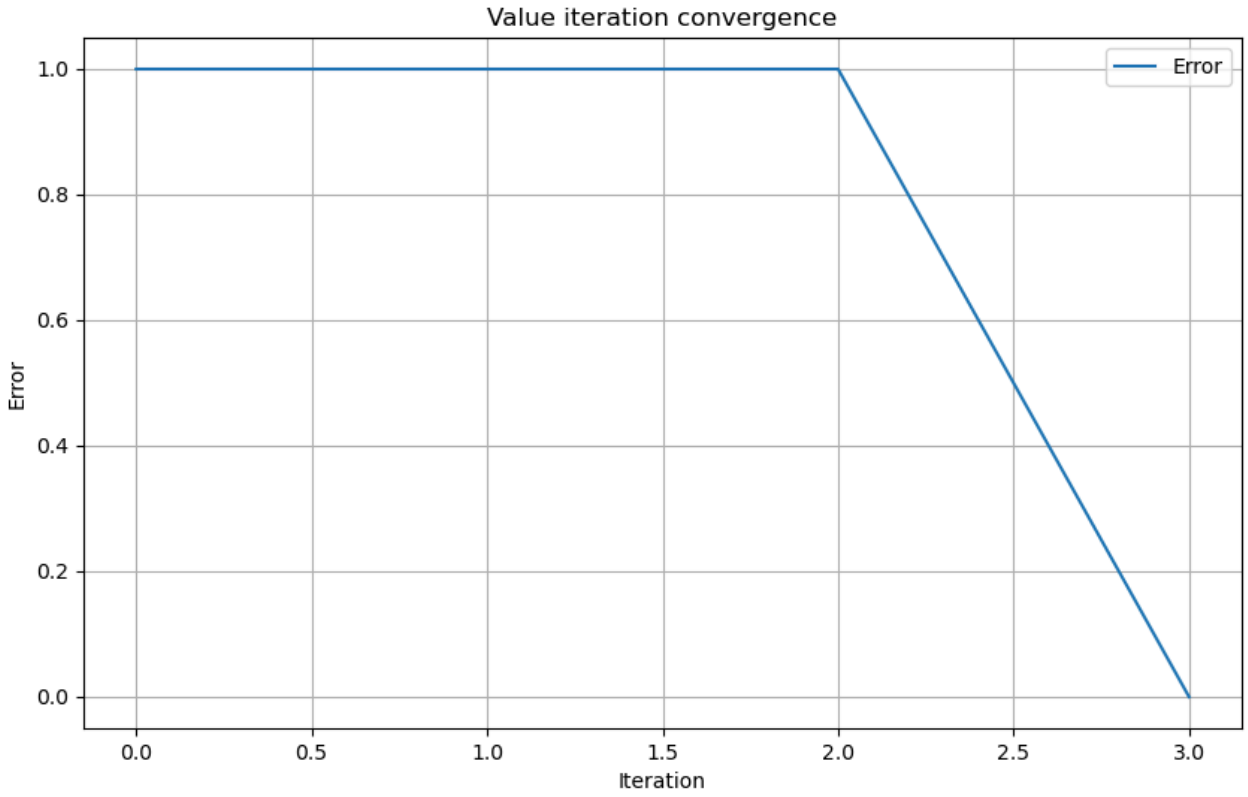
The method I chose for doing this was to go through each state and find the one that will have the best reward. The convergence message is to see whether the optimal paths changed was less than .001 because that means that there were no changes being made. I picked it because it made sense in my head (keep going until the changes are miniscule and arbitrary) and it was easy to implement. You just get the max change for each state and check if it is less than a threshold which is really easy to do

Value iteration:
Starting Value Iteration...
Finished after 4 iterations and 0.0024886131286621094 seconds
Done!
Iteration 1 values:
[[0. 0. 0. 0. 0.]
[0. 0. 0. 0. 0.]
[0. 0. 0. 0. 0.]
[0. 0. 0. 0. 0.]
[0. 0. 0. 0. 0.]]

Iteration 2 values:
[[0. 0. -1. -1. -1.]
[0. -1. -1. -1. -1.]
[-1. -1. -1. -1. -1.]
[-1. -1. -1. -1. 0.]
[-1. -1. -1. 0. 0.]]

Iteration 3 values:
[[0. 0. -1. -2. -2.]
[0. -1. -2. -2. -2.]
[-1. -2. -2. -2. -1.]
[-2. -2. -2. -1. 0.]
[-2. -2. -1. 0. 0.]]

Final values:
[[0. 0. -1. -2. -3.]
[0. -1. -2. -3. -2.]
[-1. -2. -3. -2. -1.]
[-2. -3. -2. -1. 0.]
[-3. -2. -1. 0. 0.]]



Optimal Policy:

. < < < V

^ ^ ^ ^ V

^ ^ ^ > V

^ ^ > > V

^ > > > .

The convergence method was essentially the same. Go through each state and get the max change between each possible move. The change comes into play when checking against a threshold. Instead of checking against an arbitrary small number, we just see if the change was 0. If the change is 0, then no changes were made and we have converged. Again, I picked it because it made the most sense in my head, we go until no changes were made, and it was easy to implmenet.

When running, the policy iteration was surprisingly much faster taking about .0005 seconds to converge vs .002 (though this difference could be due to implementation details). Policy iteration can be better because it is faster it was also a bit easier for me to implement compared to value iteration. According to google, it is also better when trying to getting intermediate policies and is generally more stable