Reinforcement Learning

Project Report

Class and Section: CPE471.1001

Name: Lucas Pinto

Due Date: February 12, 2025

GitHub Repository: https://github.com/PieFlavr/CPE471-Project-1

World-Agent Design

The default environment is modeled by a 5x5 “Grid World”, with an “Agent” that can do *‘up’, ‘down’, ‘right’ and ‘left’* actions. Any invalid actions such as attempting to move off the grid can be attempted by the Agent, but they do not let the Agent out. The Agent and Grid World are implemented in the `agent.py` and `grid_world.py` file respectively.

Agent Rewards

The Agent is rewarded if it reaches the goal of the Grid World, which by default is set on the bottom right of the Grid World. The agent is heavily punished for attempting to make an invalid move, and is only lightly punished for every step it takes. This was done in an effort to make the Agent attempt to follow the least number of steps to reach the goal.

By default, the reward for reaching the goal is just the area of the grid, (grid_length*grid_width). The punishments are always (-1) for every step and (-5) for attempting to move out of bounds.

From testing higher grid sizes such as 100x100 Grid Worlds, this reward structure begins to break down as the Agent loses motivation to avoid invalid moves. Additionally, the Agent seems to create additional loops during its path even during the final action sequence. It is likely due to the relative scale of the goal reward to the punishments breaking down as the grid gets larger, but increasing the training episodes accordingly such as to 100k amend the issue.

Learning Algorithms

There are only two learning algorithms currently implemented in this project, the ordinary ***Q-Learning algorithm and the Q(λ) algorithm***. While it was possible to attach the Q-Tables to the Agent (to demonstrate policy), for the scope it was easier to leave it as a variable per training run. As for the eligibility traces of Q(λ), they were implemented as a variable per-episode. While the algorithms were implemented in such a way that the action selection algorithm could be interchangeable, the only one implemented was ***Epsilon-Greedy Selection***. For the most part, all of these were implemented in the `learning.py` file.

Algorithmically, they essentially worked as follows…

Q-LEARNING UPDATE

Initialize $Q(s,a)$ arbitrarily
Repeat (*for each episode*):
..> Initialize s
..> Repeat (*for each **step** of episode*):
..> ..> Choose a from s using policy derived from Q
..> ..> Take action a, observe r, s'
..> ..> $Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma max_{a'} Q(s',a') - Q(s,a)\right]$
..> ..> $s \leftarrow s'$
..> until s is terminal

Q(λ) UPDATE

Initialize $Q(s,a)$ arbitrarily and $e(s,a) = 0$, for all s,a
Repeat (*for each episode*):
..> Initialize s,a
..> Repeat (*for each **step** of episode*):
..> ..> Take action a, observe r,s'
..> ..> Choose a' from s' using policy derived from Q
..> ..> $a^* \leftarrow arg\,max_b Q(s',a')$ (if a' ties for max, $a^* \leftarrow a'$)
..> ..> $\delta \leftarrow r + \gamma Q(s',a') - Q(s,a)$
..> ..> $e(s,a) \leftarrow e(s,a) + 1$
..> ..> For all s,a:
..> ..> ..> $Q(s,a) \leftarrow Q(s,a) + \alpha\delta e(s,a)$
..> ..> ..> If $a' = a^*$, then $e(s,a) \leftarrow \gamma e(s,a)$
..> ..> ..> else $e(s,a) \leftarrow 0$
..> ..> $s \leftarrow s'$; $a \leftarrow a'$
..> until s is terminal

Report Data Parameters

For this report, the training data for the 5x5 Grid World was produced with the following parameters:

	Q-LEARNING	**Q(λ)**
Update Parameters	$\alpha = 0.15,\ \gamma = 0.95,$ $\epsilon = 0.1$	$\alpha = 0.15,\ \gamma = 0.95,$ $\epsilon = 0.1,\ \lambda = 0.5$
Episodes	1000	1000
Agent Start Position	(0,0)	(0,0)
Goal Position	(4,4)	(4,4)
Goal Reward	25	25
Invalid Move Reward	-5	-5
Step Reward	-1	-1

Additionally, the training data for the 10x10 Grid World was produced with the same parameters except as follows:

	Q-LEARNING	**Q(λ)**
Goal Position	(9,9)	(9,9)
Goal Reward	25	25

With these parameters, it should take no less than a few seconds to generate the data required for the program to make its figures. Although, tests were performed with much higher grid sizes and episode counts. They worked although the time obviously grew with the increase of either. Anything above 10k episodes and a 100x100 grid size will begin to take more than a minute to generate everything, especially the animation of the action sequences. Additionally, although now mitigated, during programming possible issues with memory did occur so be warned of trying extremely large values.

RESULTs

The results of the above runs are shown below, but more extensively additional images and .csv data were recorded in the `project_data` folder. Furthermore, the following demo videos (5x5 and 10x10) contained example runs (not the same as analyzed data)…

- 5x5 DEMO: https://youtu.be/a9IlX-IIqPI
- 10x10 DEMO: https://youtu.be/QAEhNdgtyeM

While this report talks of the parameters of the 10x10, the figures and results are not discussed here. More detail, figures, and data are stored in the GitHub project repository: https://github.com/PieFlavr/CPE471-Project-1

5x5 Q-Learning + Q(λ) Results

With regards to the Q-Tables, both seem to evolve as expected. The steps taken per episode seem to converge around 8 steps, which is the expected behavior. With regards to converging to that number, it seems Q(λ) is faster with the onset of the behavior than Q-Learning, but appears to have a higher variation in efficiency. This is likely due to *Epsilon-Greedy Selection* as it either selects entirely randomly or exactly the best, which the *Eligibility Traces* in Q(λ)'s implementation might encourage otherwise bad behavior. You can see this in the action sequences as well, where Q-Learning has initially an almost completely random behavior, Q(λ) seems to almost prefer really long "strides". Considering the dimensionality of the Grid World being, well, a grid where diagonal movement practically doesn't exist, this is somewhat expected behavior.

5x5 Q-Learning Figures

Q-Tables

5x5 Grid World
Episodes: 100, Alpha: 0.15, Gamma: 0.95, Epsilon: 0.1
Rewards: [25, -1, -5]
Agent Start: (0, 0), Goal: (4, 4)
Trained w/ Q-Learning and Epsilon-Greedy Selection

First Q-table

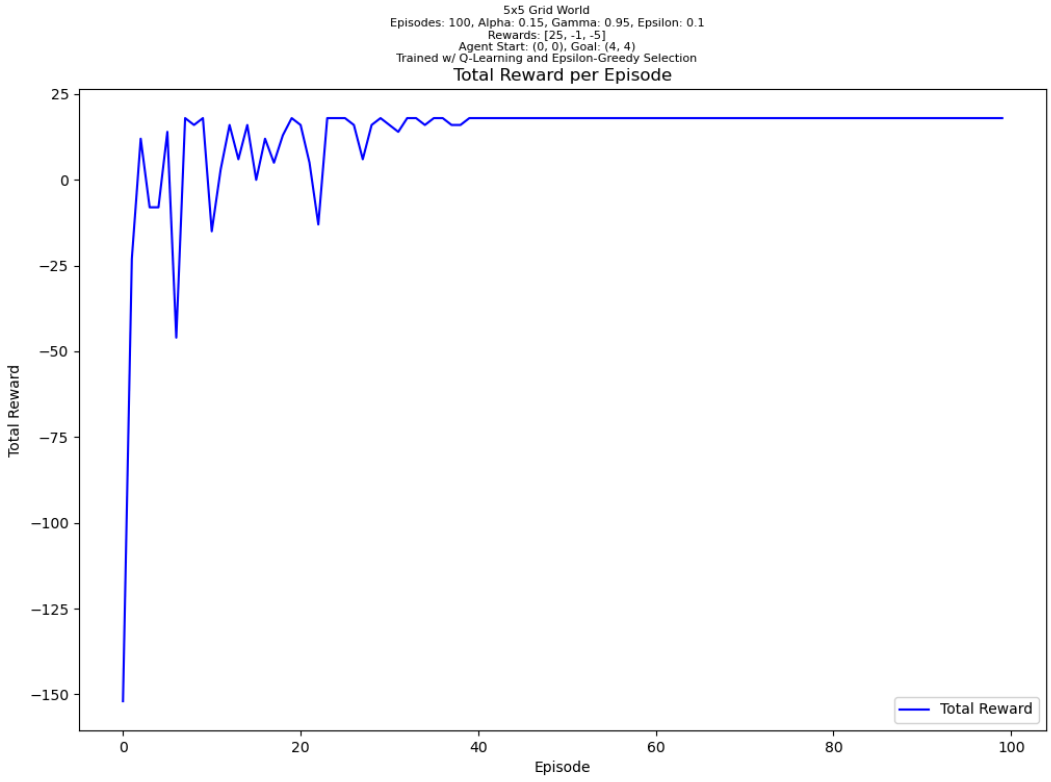
State(x,y)	up	down	left	right
(0,0)	-0.75	-0.40725	-0.75	-0.27749999999999997
(0,1)	-0.298875	-0.40725	-0.75	-0.298875
(0,2)	-0.298875	-0.27749999999999997	-0.75	-0.298875
(0,3)	-0.298875	-0.15	-0.75	-0.15
(0,4)	-0.15	0.0	0.0	0.0
(1,0)	-0.75	-0.27749999999999997	-0.3352125	-0.27749999999999997
(1,1)	-0.298875	-0.27749999999999997	-0.3352125	-0.298875
(1,2)	-0.298875	-0.27749999999999997	-0.171375	-0.298875
(1,3)	-0.298875	-0.15	-0.171375	-0.15
(1,4)	-0.15	0.0	0.0	0.0
(2,0)	-0.75	-0.27749999999999997	-0.3200896875	-0.27749999999999997
(2,1)	-0.298875	-0.27749999999999997	-0.3200896875	-0.298875
(2,2)	-0.298875	-0.27749999999999997	-0.171375	-0.298875
(2,3)	-0.298875	-0.15	-0.171375	-0.15
(2,4)	-0.15	0.0	0.0	0.0
(3,0)	-0.75	-0.27749999999999997	-0.3200896875	-0.27749999999999997
(3,1)	-0.298875	-0.27749999999999997	-0.3200896875	-0.298875
(3,2)	-0.298875	-0.27749999999999997	-0.171375	-0.15
(3,3)	-0.15	-0.15	-0.171375	-0.15
(3,4)	-0.15	0.0	0.0	0.0
(4,0)	-0.75	-0.27749999999999997	-0.3200896875	-0.75
(4,1)	-0.30192093750000004	-0.27749999999999997	-0.3200896875	-0.75
(4,2)	-0.30192093750000004	-0.15	-0.171375	-0.75
(4,3)	-0.15	3.75	0.0	0.0
(4,4)	0.0	0.0	0.0	0.0

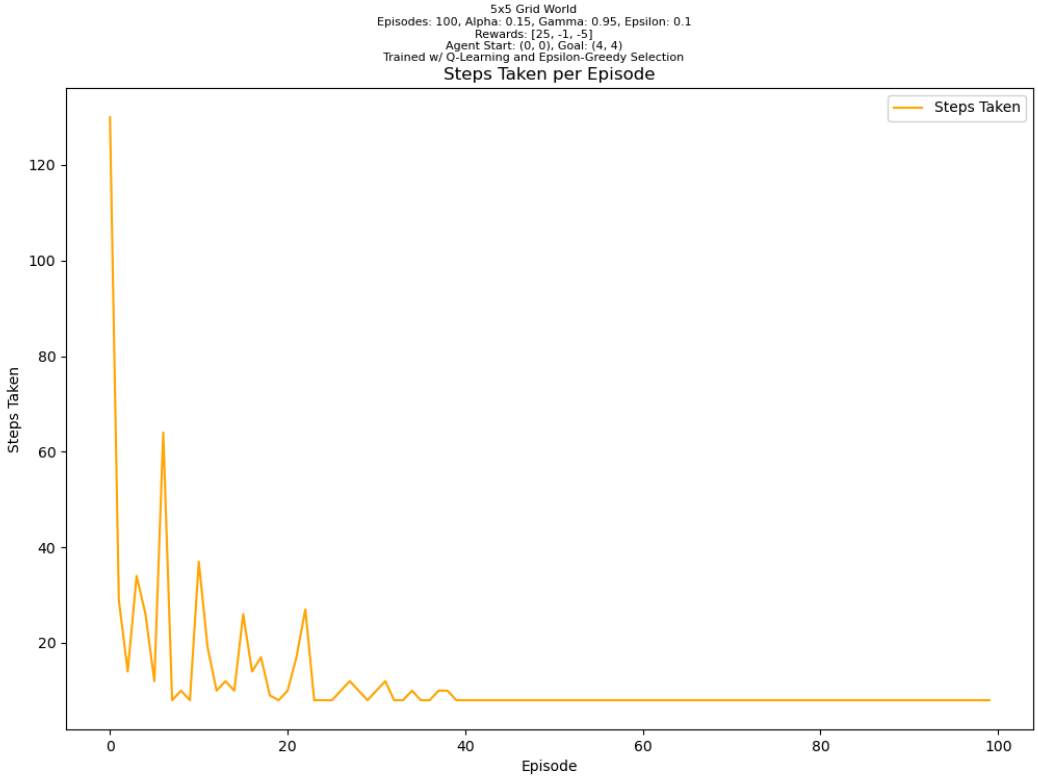
5x5 Grid World
Episodes: 100, Alpha: 0.15, Gamma: 0.95, Epsilon: 0.1
Rewards: [25, -1, -5]
Agent Start: (0, 0), Goal: (4, 4)
Trained w/ Q-Learning and Epsilon-Greedy Selection

Last Q-table

State(x,y)	up	down	left	right
(0,0)	-2.96641843359375	-2.7960639664660616	-2.96641843359375	10.644843504437159
(0,1)	-2.2284870655170157	-2.087864329824768	-2.2331671875	-1.9338434329403174
(0,2)	-1.588020512060701	-1.4924707816862999	-1.494375	-0.8966472358235018
(0,3)	-1.1714591405157027	-1.1654731915087293	-1.494375	-1.0719349293574265
(0,4)	-0.9322440896783968	-1.494375	-1.494375	-0.9113874741024756
(1,0)	-2.2331671875	12.731676241396373	-2.1109972748452135	-2.098867938245755
(1,1)	-1.6241708035113382	14.689632037341433	-1.7044924058030382	-1.5840386946031093
(1,2)	-1.162095533817085	-1.0907521478089324	-1.162980512034958	16.612359804115343
(1,3)	-0.8037713315934576	-0.6792935112890626	-0.8416189028555848	-0.47938623253713697
(1,4)	-0.61521336796875	-0.75	-0.6605331215683593	0.11170214273798196
(2,0)	-1.494375	-1.298731692643969	-1.6256541119479462	-1.435704580007809
(2,1)	-1.1036520075714253	-1.032182148144598	-1.1897320010692505	0.9838662214253725
(2,2)	-0.7883413265014878	-0.6532766361328124	-0.6621985694229052	18.57186305133245
(2,3)	-0.46251092109375	-0.4102959375	-0.3352125	2.1886480660136796
(2,4)	-0.298875	-0.75	-0.33825843749999995	6.104962986307775
(3,0)	-1.494375	-0.3106148468237345	-1.0198436017733223	-0.9353313744296922
(3,1)	-0.7581519712674034	7.6701035372420305	-0.6492828328651904	-0.6633802666523438
(3,2)	-0.46725496875000005	-0.40725	-0.3352125	20.61061639835905
(3,3)	-0.298875	9.589153443126671	-0.171375	-0.15
(3,4)	-0.15	-0.75	-0.171375	21.443956071595693
(4,0)	-0.75	-0.45521470526597047	-0.6529167706313243	-0.75
(4,1)	-0.45224557734375004	2.9553770393660823	-0.46768901484375003	-0.75
(4,2)	-0.30192093750000004	22.749759187631923	-0.171375	-0.75
(4,3)	-0.15	24.99998462534173	0.0	0.0
(4,4)	0.0	0.0	0.0	0.0

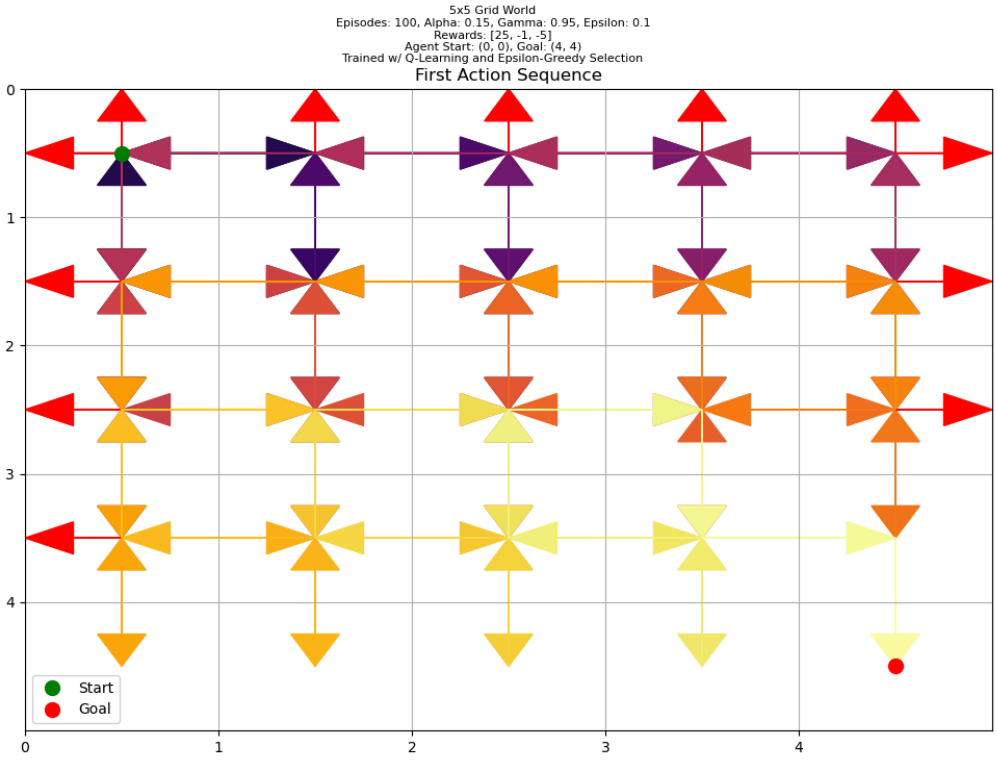
Reward and Step Plots

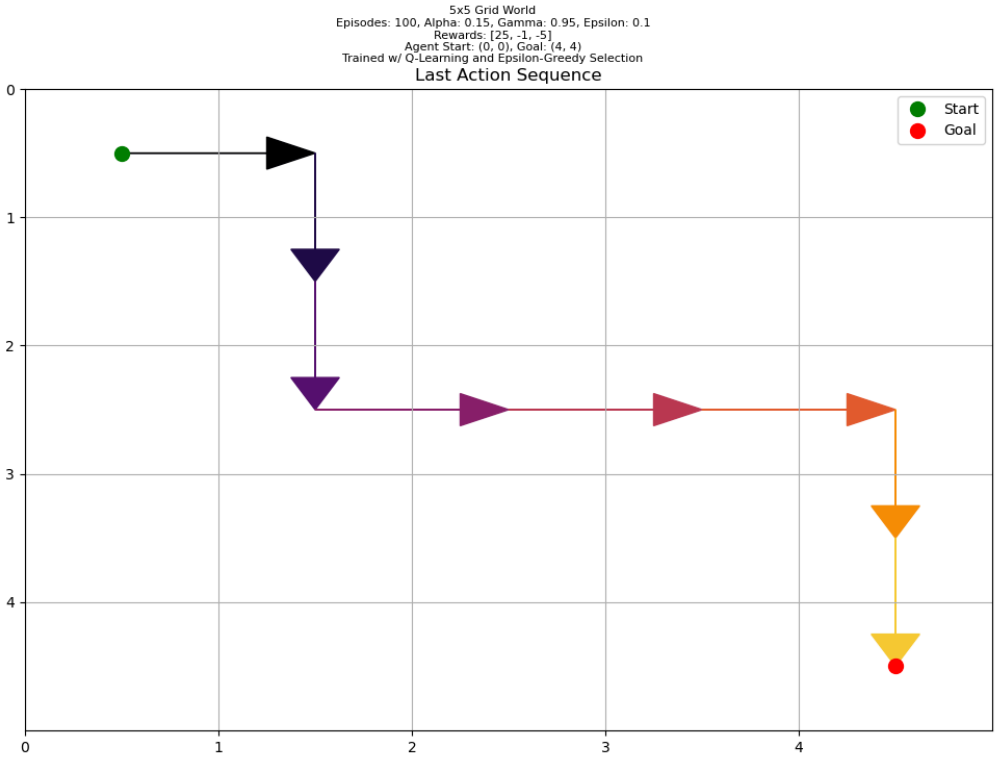




Action Sequences

(Darker = Earlier Actions ; Brighter = Later Actions; Red = Invalid Steps)





5x5 Q(λ) Figures

Q-Tables

5x5 Grid World
Episodes: 100, Alpha: 0.15, Gamma: 0.95, Epsilon: 0.1
Rewards: [25, -1, -5]
Agent Start: (0, 0), Goal: (4, 4)
Trained w/ Q-Lambda and Epsilon-Greedy Selection

First Q-table

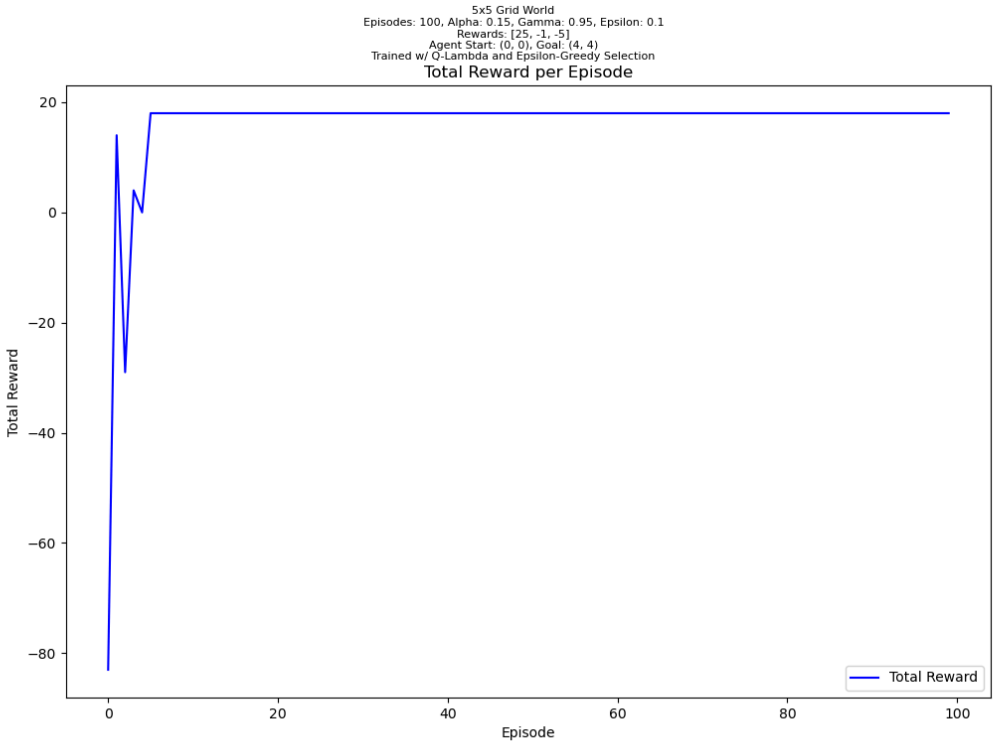
State(x,y)	up	down	left	right
(0,0)	-0.9646881632202883	-0.7345644607324145	-1.0226042347424475	-0.5739036520893623
(0,1)	-0.635737011502662	-0.42163324920479256	-0.8881252042317673	-0.599680501133483
(0,2)	-0.5718594720100892	-0.42125990504914723	-0.886470493292619	-0.2873063016686717
(0,3)	-0.5710734843139939	-0.4211444435340163	-0.8859587525053347	-0.28622895264281006
(0,4)	-0.5708304074400338	0.0	0.0	0.0
(1,0)	-0.8924287412407627	-0.299849981559501	-0.3483655692387846	0.0
(1,1)	-0.31547364538842276	-0.5445646240099395	-0.3330738931816374	-0.4213692539762521
(1,2)	-0.30821009926127785	-0.30024096388643917	-0.35009845489834507	-0.5640904740392482
(1,3)	-0.316296766076714	-0.3002285896593721	-0.35004361067865775	-0.2867977950374949
(1,4)	-0.3162707150723625	0.0	0.0	0.0
(2,0)	-0.8869551422770162	-0.28832661532003473	0.0	0.0
(2,1)	-0.5713036925815826	-0.2912139269895465	-0.3100894271004828	0.0
(2,2)	-0.29729247787272944	-0.5353102052300593	-0.31698753287497133	-0.3500723661904745
(2,3)	-0.3005690781156114	-0.28726981088240056	-0.2926085800882018	-0.541990451891607
(2,4)	-0.28898907554189596	0.0	0.0	0.0
(3,0)	-0.8862270646144269	-0.2867938202408995	0.0	0.0
(3,1)	-0.5709578556918528	-0.28798698998084105	0.0	0.0
(3,2)	-0.42120498145363017	-0.29049892627545465	-0.3069204488662802	0.0
(3,3)	-0.2957872132114832	-0.2905165431024705	-0.3069985289860183	-0.3057997727768326
(3,4)	-0.29582430126835874	0.0	0.0	0.0
(4,0)	-0.68026962890625	0.14680078124999996	0.0	0.0
(4,1)	-0.47312807373046867	0.6248437499999999	0.0	0.0
(4,2)	-0.37473583502197266	1.63125	0.0	0.0
(4,3)	-0.32799952163543705	3.75	0.0	0.0
(4,4)	0.0	0.0	0.0	0.0

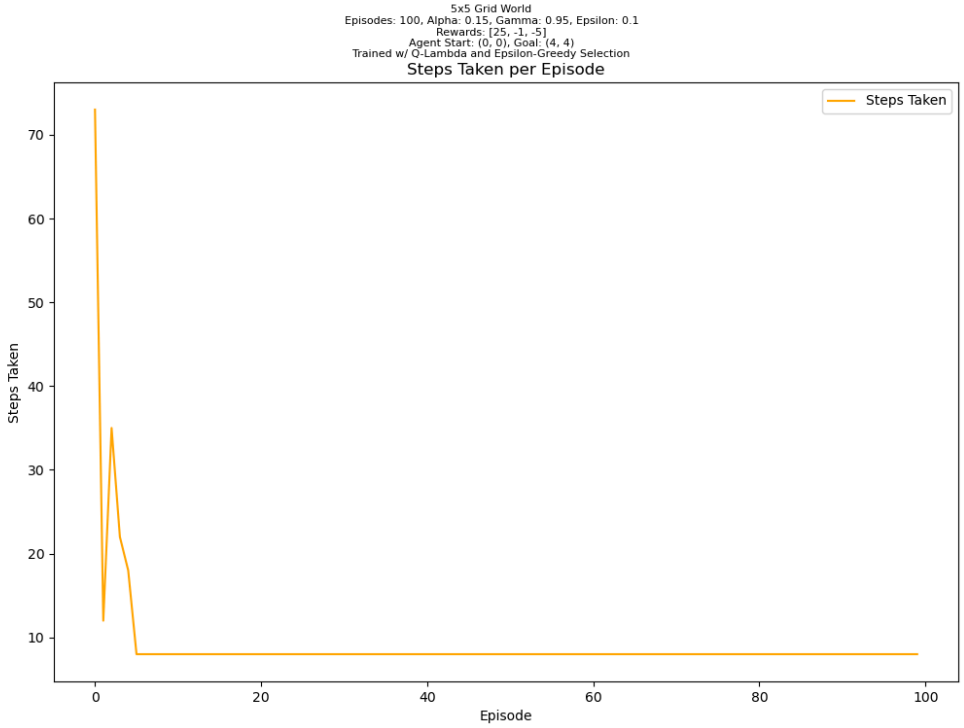
5x5 Grid World
Episodes: 100, Alpha: 0.15, Gamma: 0.95, Epsilon: 0.1
Rewards: [25, -1, -5]
Agent Start: (0, 0), Goal: (4, 4)
Trained w/ Q-Lambda and Epsilon-Greedy Selection

Last Q-table

State(x,y)	up	down	left	right
(0,0)	-1.8883869842444265	-1.5188455067953834	-1.865522251060926	11.412483974097672
(0,1)	-1.236483620979339	-0.9589142938750213	-0.8881252042317673	-1.0155214182843288
(0,2)	-0.8498813321321032	-0.68781542844452	-0.886470493292619	-0.6054089325430899
(0,3)	-0.5710734843139939	-0.6905220757740717	-0.8859587525053347	-0.5787697598031143
(0,4)	-0.5708304074400338	-1.1856810672796867	-0.917223299536183	-0.6307486493589172
(1,0)	-0.8924287412407627	-0.847133140503967	-0.9784361763554771	13.072850266308455
(1,1)	-0.7949527704748477	-0.7977300651090511	-0.6173931750939786	-0.606551298542717
(1,2)	-0.5908777560543625	-0.6547972560357049	-0.6342270967458988	-0.5640904740392482
(1,3)	-0.5970393272866541	-0.5837304673861826	-0.6355045275944825	-0.7262162180677869
(1,4)	-0.5972551135521614	-1.0887692908049793	-0.7131985069578513	-0.6798764033908403
(2,0)	-0.8869551422770162	14.81732363772508	-0.5985389965228115	-0.36909796085632324
(2,1)	-0.5713036925815826	-0.2912139269895465	-0.3100894271004828	16.651663165987102
(2,2)	-0.32279864795099666	-0.5353102052300593	-0.31698753287497133	-0.3500723661904745
(2,3)	-0.45996332528231065	-0.9009585355880304	-0.5761632590216669	-0.541990451891607
(2,4)	-0.7547653839972467	-1.0669035712378738	-0.6671654131323651	-0.5797187884991568
(3,0)	-0.8862270646144269	-0.2867938202408995	-0.22473964826660164	0.006195585937499942
(3,1)	-0.5709578556918528	-0.28798698998084105	-0.12027670986328128	18.581496395550573
(3,2)	-0.42120498145363017	-0.29049892627545465	-0.3069204488662802	0.0
(3,3)	-0.2957872132114832	-0.2905165431024705	-0.3069985289860183	-0.3057997727768326
(3,4)	-0.29582430126835874	-0.9046711336824355	-0.3256234393314433	0.0
(4,0)	-0.68026962890625	0.43159335937499993	0.0	0.0
(4,1)	-0.47312807373046867	20.61239674539086	0.0	0.0
(4,2)	-0.37473583502197266	22.7499796783823	0.0	0.0
(4,3)	-0.32799952163543705	24.999997813081592	0.0	0.0
(4,4)	0.0	0.0	0.0	0.0

Reward and Step Plots





Action Sequences

(Darker = Earlier Actions ; Brighter = Later Actions; Red = Invalid Steps)