INF8953DE Assignment 3

Emile Dimas

TOTAL POINTS

94 / 100

QUESTION 1

Qn 1.1 REINFORCE with episodal returns 25 pts

1.1 Qn 1.1.a: Implement REINFORCEv1 agent **20 / 20**

✓ **- 0 pts** Correct

1.2 Qn 1.1.b: Why can we allow ourselves to use γ=1.0 here? **3 / 3**

✓ **- 0 pts** Correct

1.3 Qn 1.1.c: If you have implemented everything correctly, you will notice that training iterations tend to take a bit longer towards the end compared to early stages of the training, why? **0 / 2**

✓ **- 2 pts** this is due to the fact that episodes tend to be longer towards the end of the training phase (better policies)

QUESTION 2

Qn 1.2 REINFORCE with returns 15 pts

2.1 Qn 1.2.a Implement REINFORCEv2 agent **10 / 10**

✓ **- 0 pts** Correct

2.2 Qn 1.2.b: Plot and compare the performance of the REINFORCEv1 and REINFORCEv2 agents for γ=1 . **5 / 5**

✓ **- 0 pts** Correct

QUESTION 3

Qn 1.3 REINFORCE WITH baseline 25 pts

3.1 Qn 1.3.a Implement 'REINFORCEv2+B' agent **15 / 15**

✓ **- 0 pts** Correct

3.2 Qn 1.3.b : Does introducing baselines have a meaning beyond variance reduction? **3 / 5**

✓ **- 0 pts** Correct

✓ **- 2 pts** you should have mentioned something along the lines of the fact that the Q(s, a) - V(s) acts as an advantage estimator, A(s, a). Basically you need t mention that some states have higher state values by default and baseline is there to reduce this undesired effect.

3.3 Qn 1.3.c Plot and compare REINFORCEv2+B for $\gamma \in \{0.95, 0.975, 0.99, 0.995, 1\}$ **5 / 5**

✓ **- 0 pts** Correct

QUESTION 4

Actor Critic 35 pts

4.1 Qn 2.1 Implement a one-step Actor-Critic agent **15 / 15**

✓ **- 0 pts** correct

4.2 Qn 2.2: Eventhough the previous REINFORCEv2+B agent used a value estimator network similar to that of the Actor-Critic agent why is not called an Actor-Critic method ? **2 / 3**

✓ - 1 pts insufficient explanation

4.3 Qn 2.3: How does the Actor-Critic algorithm reduces variance? What about bias? We are using one-step rewards here, is there a way we can strike a balance between variance and bias? 5 / 5

✓ - 0 pts Correct

4.4 Qn 2.4: Challenge! Can you tweak the hyperparameters of Actor-Critic to achieve better performance? Compare your results againts what you already have in section 3.1, in a single plot. 5 / 5

✓ - 0 pts Correct

4.5 Qn 2.5: Compare and plot `REINFORCEv2+B' method and 'ACTOR-CRITIC' method 4 / 5

✓ - 0 pts Correct

✓ - 1 pts Poor AC performance

4.6 Qn 2.6: Plot all methods 2 / 2

✓ - 0 pts Correct

```
        show_video("./gym-results")
        print(f'Reward: {reward_episode}')
```

4 Qn 1. REINFORCE ALGORITHM [65 Marks]

4.1 Qn 1.1 REINFORCE with episodal returns [25 Marks]

4.1.1 Qn1.1.a: Implement a `REINFORCEv1` agent [20 Marks]

Implement a REINFORCE agent below with the following policy gradient computation.

$\nabla_\theta J(\theta) = \sum_j \sum_t G_0^j \nabla_\theta ln \pi_\theta(a_t^j | s_t^j)$ \ where $G_0^j = \sum_{k=0}^{\infty} \gamma^k R_{k+1}^j$ is the discounted return for the start state, s_0^j for the episode j. . \

Note that this is different from the REINFORCE algorithm we have seen in the class since we are using only the episodal return in the policy gradient computation for all state updates instead of using the corresponding return from individual states. We will implement the in-class version of REINFORCE in the next part.

You will be graded primarily on the output of the agent.train() and agent.evaluate() functions for this question.

[ ]:
```
# Insert your code and run this cell
class REINFORCEv1Agent(BaseAgent):
    """ REINFORCE agent with total trajectory reward.
    """

    def optimize_model(self, n_episodes: int):
        """ YOU NEED TO IMPLEMENT THIS METHOD

            This method is called at each training iteration and is responsible␣
↪for
            (i) gathering a dataset of episodes
            (ii) computing the expectation of the policy gradient.
                 Note that you will only be computing the loss value

            HINTS:

                * Note that policy network model (self.policy_model) outputs␣
↪the
                probability of taking each discrete action. Hence, you need
                to sample from this distribution. Take a look at `self.
↪evaluate()`
                method in the `BaseAgent` class.

                * Keep in mind that policy network takes batches of states as
                input, as opposed to a single state vector. This is by design,
                and good/common practice, however, you need to keep an eye on
```

```
            the input/output dimensions.

        """

        # ============================================================

        # INSERT YOUR CODE HERE !
        loss = torch.tensor([0.0], requires_grad=True).to(self.device)
        total_rewards = np.empty(n_episodes)
        for episode in range(n_episodes):
          nested_loss = torch.tensor([0.0], requires_grad=True).to(self.device)
          states = []
          rewards = []
          actions = []

          observation = self.monitor_env.reset()
          done = False

          while not done:
            states.append(observation)
            observation = torch.tensor(observation, dtype=torch.float)[None, :].
↪to(self.device)
            probs = self.policy_model.forward(observation)
            action = torch.multinomial(probs, 1)[0] # draw samples from dist
            nested_loss = nested_loss + torch.log(probs[0,int(action)])
            actions.append(action.detach().cpu().numpy())
            observation, reward, done, info = self.monitor_env.step(int(action))
            rewards.append(reward)


          total_rewards[episode] = sum(rewards)


          Gs = [r*self.gamma**i for i,r in enumerate(rewards)]
          G0 = sum(Gs)
          loss =  - G0*nested_loss + loss
          self.monitor_env.close()

        # ============================================================

        self.policy_optimizer.zero_grad()
        loss.backward()
        self.policy_optimizer.step()
        return total_rewards
```

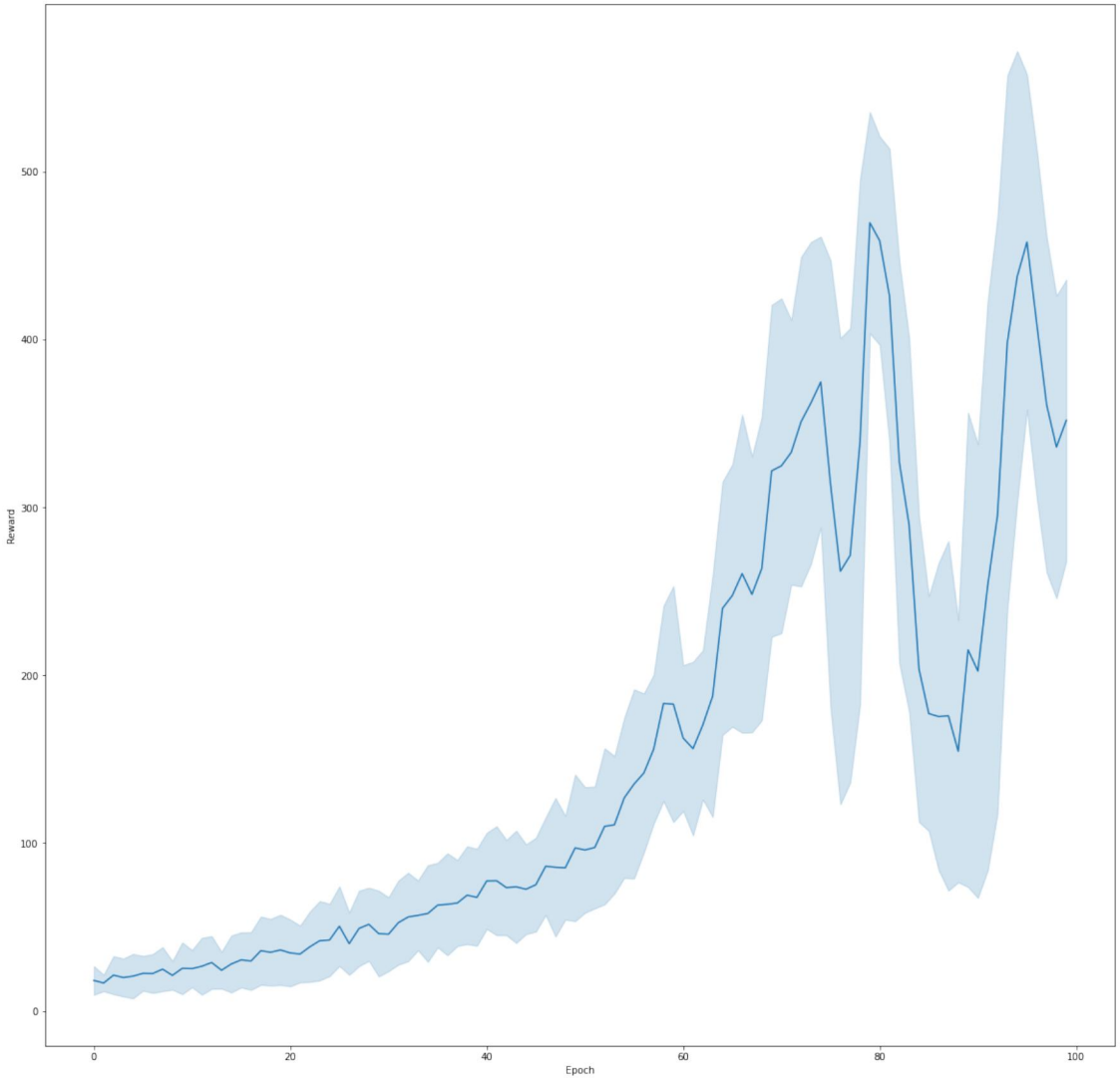
[ ]:
```
# You will be graded on this output this cell, so kindly run it
```

```python
# This is an example configuration that is tuned for the above question.
# keep the same config
config = {
    'env_id': 'CartPole-v1',
    'seed': 8953,
    'gamma': 1.0,
    'policy_layers': [16, 8],
    'policy_learning_rate': 1e-2,
    'use_baseline': False,
}
agent = REINFORCEv1Agent(config)
REINFORCEv1_rewards = agent.train(n_episodes=50, n_iterations=100)
```

```
the device is: cpu
Iteration 1/100: rewards 18.28 +/- 8.44
Iteration 2/100: rewards 16.74 +/- 4.84
Iteration 3/100: rewards 21.42 +/- 11.21
Iteration 4/100: rewards 20.02 +/- 11.18
Iteration 5/100: rewards 20.84 +/- 13.11
Iteration 6/100: rewards 22.5 +/- 10.26
Iteration 7/100: rewards 22.4 +/- 11.46
Iteration 8/100: rewards 24.98 +/- 12.99
Iteration 9/100: rewards 21.3 +/- 8.47
Iteration 10/100: rewards 25.46 +/- 15.24
Iteration 11/100: rewards 25.3 +/- 10.97
Iteration 12/100: rewards 26.72 +/- 16.78
Iteration 13/100: rewards 28.94 +/- 15.54
Iteration 14/100: rewards 24.36 +/- 10.87
Iteration 15/100: rewards 28.08 +/- 16.78
Iteration 16/100: rewards 30.5 +/- 16.24
Iteration 17/100: rewards 29.8 +/- 17.04
Iteration 18/100: rewards 35.96 +/- 20.14
Iteration 19/100: rewards 35.06 +/- 19.67
Iteration 20/100: rewards 36.44 +/- 20.72
Iteration 21/100: rewards 34.64 +/- 19.68
Iteration 22/100: rewards 34.02 +/- 16.76
Iteration 23/100: rewards 38.34 +/- 20.77
Iteration 24/100: rewards 41.92 +/- 23.4
Iteration 25/100: rewards 42.36 +/- 21.33
Iteration 26/100: rewards 50.56 +/- 23.49
Iteration 27/100: rewards 40.18 +/- 18.39
Iteration 28/100: rewards 49.24 +/- 22.24
Iteration 29/100: rewards 51.7 +/- 21.54
Iteration 30/100: rewards 46.14 +/- 25.28
Iteration 31/100: rewards 45.78 +/- 21.8
Iteration 32/100: rewards 52.66 +/- 24.78
Iteration 33/100: rewards 56.06 +/- 26.08
Iteration 34/100: rewards 57.02 +/- 20.59
```

```
Iteration 35/100: rewards 58.14 +/- 28.47
Iteration 36/100: rewards 63.1 +/- 24.95
Iteration 37/100: rewards 63.64 +/- 30.06
Iteration 38/100: rewards 64.36 +/- 25.4
Iteration 39/100: rewards 69.04 +/- 28.83
Iteration 40/100: rewards 67.76 +/- 28.6
Iteration 41/100: rewards 77.56 +/- 28.36
Iteration 42/100: rewards 77.66 +/- 32.13
Iteration 43/100: rewards 73.54 +/- 28.08
Iteration 44/100: rewards 74.02 +/- 33.1
Iteration 45/100: rewards 72.58 +/- 26.48
Iteration 46/100: rewards 75.28 +/- 27.66
Iteration 47/100: rewards 86.28 +/- 28.8
Iteration 48/100: rewards 85.68 +/- 40.85
Iteration 49/100: rewards 85.36 +/- 30.6
Iteration 50/100: rewards 97.22 +/- 43.22
Iteration 51/100: rewards 95.96 +/- 37.03
Iteration 52/100: rewards 97.46 +/- 35.93
Iteration 53/100: rewards 110.06 +/- 46.01
Iteration 54/100: rewards 110.96 +/- 40.57
Iteration 55/100: rewards 127.0 +/- 47.2
Iteration 56/100: rewards 135.24 +/- 55.73
Iteration 57/100: rewards 141.84 +/- 46.89
Iteration 58/100: rewards 156.06 +/- 43.91
Iteration 59/100: rewards 183.26 +/- 57.63
Iteration 60/100: rewards 182.82 +/- 69.51
Iteration 61/100: rewards 162.62 +/- 42.99
Iteration 62/100: rewards 156.32 +/- 51.14
Iteration 63/100: rewards 170.44 +/- 44.01
Iteration 64/100: rewards 187.54 +/- 71.18
Iteration 65/100: rewards 239.84 +/- 74.6
Iteration 66/100: rewards 247.48 +/- 77.32
Iteration 67/100: rewards 260.54 +/- 93.76
Iteration 68/100: rewards 248.18 +/- 81.3
Iteration 69/100: rewards 263.62 +/- 89.5
Iteration 70/100: rewards 321.72 +/- 97.85
Iteration 71/100: rewards 324.76 +/- 98.65
Iteration 72/100: rewards 332.78 +/- 77.99
Iteration 73/100: rewards 351.02 +/- 97.15
Iteration 74/100: rewards 362.14 +/- 94.89
Iteration 75/100: rewards 374.66 +/- 85.68
Iteration 76/100: rewards 313.64 +/- 132.07
Iteration 77/100: rewards 261.96 +/- 137.43
Iteration 78/100: rewards 271.44 +/- 133.95
Iteration 79/100: rewards 339.1 +/- 154.8
Iteration 80/100: rewards 469.5 +/- 65.22
Iteration 81/100: rewards 458.82 +/- 61.44
Iteration 82/100: rewards 425.88 +/- 86.91
```

```
Iteration 83/100: rewards 326.98 +/- 118.65
Iteration 84/100: rewards 289.5 +/- 110.14
Iteration 85/100: rewards 203.74 +/- 90.21
Iteration 86/100: rewards 177.2 +/- 69.2
Iteration 87/100: rewards 175.42 +/- 90.49
Iteration 88/100: rewards 175.86 +/- 103.06
Iteration 89/100: rewards 154.72 +/- 77.21
Iteration 90/100: rewards 215.14 +/- 139.81
Iteration 91/100: rewards 202.52 +/- 133.75
Iteration 92/100: rewards 253.36 +/- 167.89
Iteration 93/100: rewards 295.36 +/- 175.84
Iteration 94/100: rewards 398.08 +/- 157.46
Iteration 95/100: rewards 437.36 +/- 133.02
Iteration 96/100: rewards 458.02 +/- 98.85
Iteration 97/100: rewards 409.1 +/- 102.56
Iteration 98/100: rewards 361.08 +/- 98.87
Iteration 99/100: rewards 335.92 +/- 89.16
Iteration 100/100: rewards 351.84 +/- 82.92
```



[ ]:
```
# You will be graded on this output this cell, so kindly run it
agent.evaluate()
```

```
<IPython.core.display.HTML object>
```

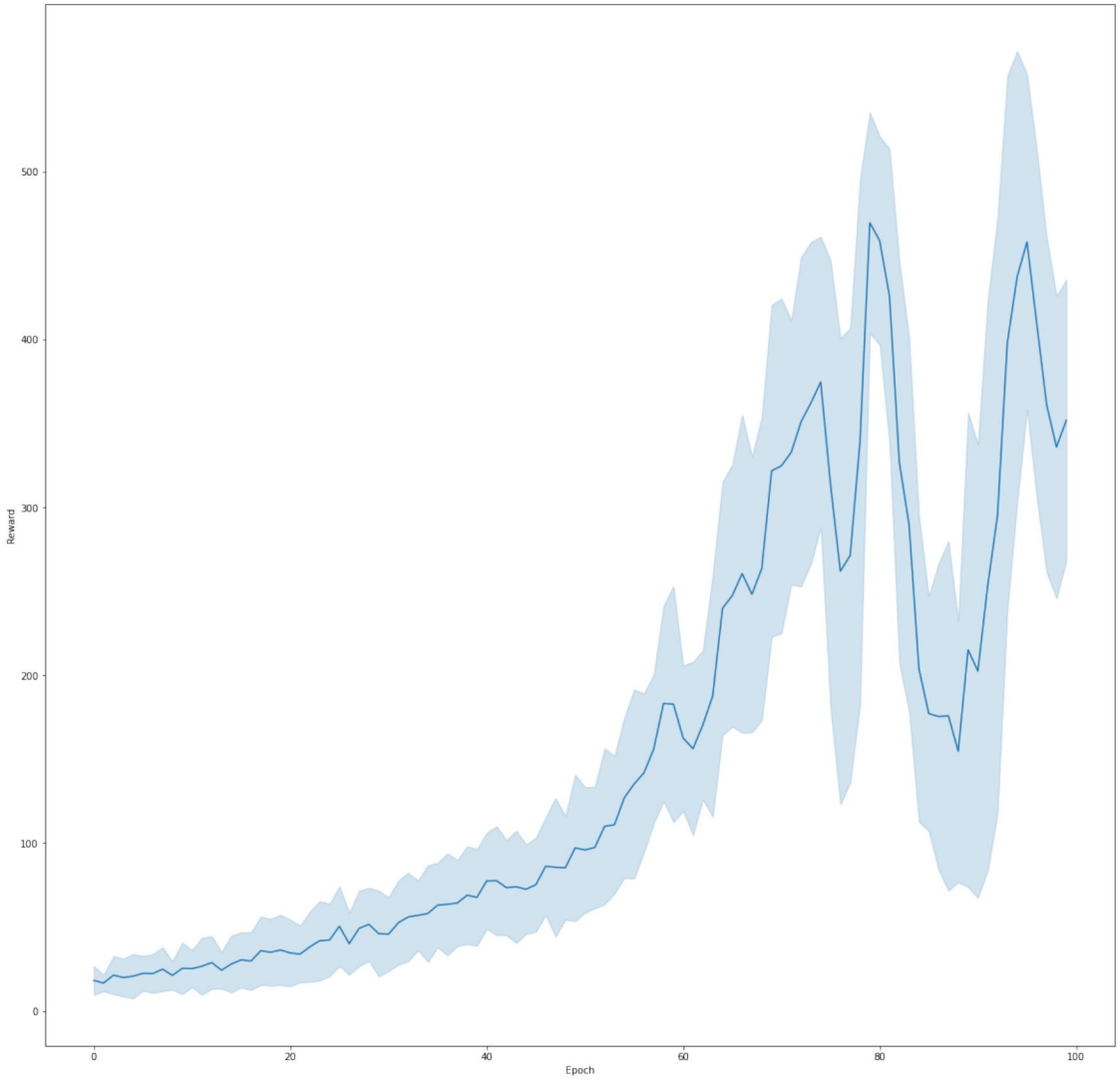
```
Reward: 246.0
```

4.1.2 Qn:1.1.b:Why can we allow ourselves to use $\gamma = 1.0$ here? [3 Marks]

We are allowed to use a $\gamma = 1$ because the environment is episodic. That means it ends after a certain number of transitions. (The reward is not inifinite)

1.1 Qn 1.1.a: Implement REINFORCEv1 agent **20 / 20**

✓ **- 0 pts** **Correct**



[ ]:
```
# You will be graded on this output this cell, so kindly run it
agent.evaluate()
```

```
<IPython.core.display.HTML object>
```

```
Reward: 246.0
```

4.1.2 Qn:1.1.b:Why can we allow ourselves to use $\gamma = 1.0$ here? [3 Marks]

We are allowed to use a $\gamma = 1$ because the environment is episodic. That means it ends after a certain number of transitions. (The reward is not inifinite)

1.2 Qn 1.1.b: Why can we allow ourselves to use $\beta=1.0$ here? **3 / 3**

✓ **- 0 pts** **Correct**

4.1.3 Qn 1.1.c: If you have implemented everything correctly, you will notice that training iterations tend to take a bit longer towards the end compared to early stages of the training, why? [2 Marks]

$\theta_{t+1} = \theta_t + \alpha G_t \frac{\nabla \pi(A_t|S_t,\theta_t)}{\pi(A_t|S_t,\theta_t)}$

The formula of G_t has a direct effect on the learning rate in the REINFORCE algorithm. (changing G_t in the formula of Reinforce update (above) has a direct effect on the learning rate and the rate of convergence), since G_t is constant in our case, the algorithm will take more time to converge towards the end, where a big learning rate is not optimal

4.2 Qn 1.2 REINFORCE with returns [15 Marks]

4.2.1 Qn 1.2.a Implement `REINFORCEv2` agent as described below. [10 Marks]

Implement a REINFORCE agent below with the following policy gradient computation.

$\nabla_\theta J(\theta) = \sum_j \sum_t G_t^j \nabla_\theta ln \pi_\theta(a_t^j|s_t^j)$ \ where $G_t^j = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}^j$ is the discounted return computed starting from the current state, s_t^j for the episode j. \

Let's call this agent REINFORCEv2.

Note that you will be graded primarily on the output of the agent.train() and agent.evaluate() functions for this question.

[ ]:
```
# Insert your code and run this cell
class REINFORCEv2Agent(BaseAgent):
    """ Not Vanilla REINFORCE Agent:
        *Not* vanilla, in the sense that we are now going to weight the action
        logprobs, proportionate to the onward return as opposed to the total
        episodic return.
    """

    def optimize_model(self, n_episodes: int):
        """ YOU NEED TO IMPLEMENT THIS METHOD

            This method is called at each training iteration and is responsible␣
 ↪for
            (i) gathering a dataset of episodes
            (ii) computing the expectation of the policy gradient.
                 Note that you will only be computing the loss value

            HINTS:
                Hints from the previous section hold here except/plus that:

                * You probably DO need to call the `BaseAgent._make_returns`
                method in this part.

                * You basically need to copy a lot of stuff you've done in the
                previous part, but have to scale the logprobs with different
```

1.3 Qn 1.1.c: If you have implemented everything correctly, you will notice that training iterations tend to take a bit longer towards the end compared to early stages of the training, why? **0 / 2**

✓ **- 2 pts** **this is due to the fact that episodes tend to be longer towards the end of the training phase (better policies)**

4.1.3 Qn 1.1.c: If you have implemented everything correctly, you will notice that training iterations tend to take a bit longer towards the end compared to early stages of the training, why? [2 Marks]

$\theta_{t+1} = \theta_t + \alpha G_t \frac{\nabla \pi(A_t|S_t,\theta_t)}{\pi(A_t|S_t,\theta_t)}$

The formula of G_t has a direct effect on the learning rate in the REINFORCE algorithm. (changing G_t in the formula of Reinforce update (above) has a direct effect on the learning rate and the rate of convergence), since G_t is constant in our case, the algorithm will take more time to converge towards the end, where a big learning rate is not optimal

4.2 Qn 1.2 REINFORCE with returns [15 Marks]

4.2.1 Qn 1.2.a Implement `REINFORCEv2` agent as described below. [10 Marks]

Implement a REINFORCE agent below with the following policy gradient computation.

$\nabla_\theta J(\theta) = \sum_j \sum_t G_t^j \nabla_\theta ln \pi_\theta(a_t^j|s_t^j)$ \ where $G_t^j = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}^j$ is the discounted return computed starting from the current state, s_t^j for the episode j. \

Let's call this agent REINFORCEv2.

Note that you will be graded primarily on the output of the agent.train() and agent.evaluate() functions for this question.

[ ]:
```
# Insert your code and run this cell
class REINFORCEv2Agent(BaseAgent):
    """ Not Vanilla REINFORCE Agent:
        *Not* vanilla, in the sense that we are now going to weight the action
        logprobs, proportionate to the onward return as opposed to the total
        episodic return.
    """

    def optimize_model(self, n_episodes: int):
        """ YOU NEED TO IMPLEMENT THIS METHOD

            This method is called at each training iteration and is responsible␣
 ↪for
            (i) gathering a dataset of episodes
            (ii) computing the expectation of the policy gradient.
                 Note that you will only be computing the loss value

            HINTS:
                Hints from the previous section hold here except/plus that:

                * You probably DO need to call the `BaseAgent._make_returns`
                method in this part.

                * You basically need to copy a lot of stuff you've done in the
                previous part, but have to scale the logprobs with different
```

```
            values.
        """

        # ============================================================================

        # INSERT YOUR CODE HERE !
        loss = torch.tensor([0.0], requires_grad=True).to(self.device)
        total_rewards = np.empty(n_episodes)
        for episode in range(n_episodes):
          nested_loss = torch.tensor([0.0], requires_grad=True).to(self.device)
          states = []
          rewards = []
          actions = []
          sub_probs = []

          observation = self.monitor_env.reset()
          done = False

          while not done:
            states.append(observation)
            observation = torch.tensor(observation, dtype=torch.float)[None, :].
↪to(self.device)
            probs = self.policy_model.forward(observation)
            action = torch.multinomial(probs, 1)[0] # draw samples from dist
            sub_probs.append(torch.log(probs[0,int(action)]))
            actions.append(action.detach().cpu().numpy())
            observation, reward, done, info = self.monitor_env.step(int(action))
            rewards.append(reward)


          total_rewards[episode] = sum(rewards)


          Gs = [r*self.gamma**i for i,r in enumerate(rewards)]
          Gk = np.cumsum(Gs[::-1])[::-1]

          for i in range(len(sub_probs)):

            loss =  - Gk[i]*sub_probs[i] + loss

          self.monitor_env.close()

        # ============================================================================



        self.policy_optimizer.zero_grad()
```

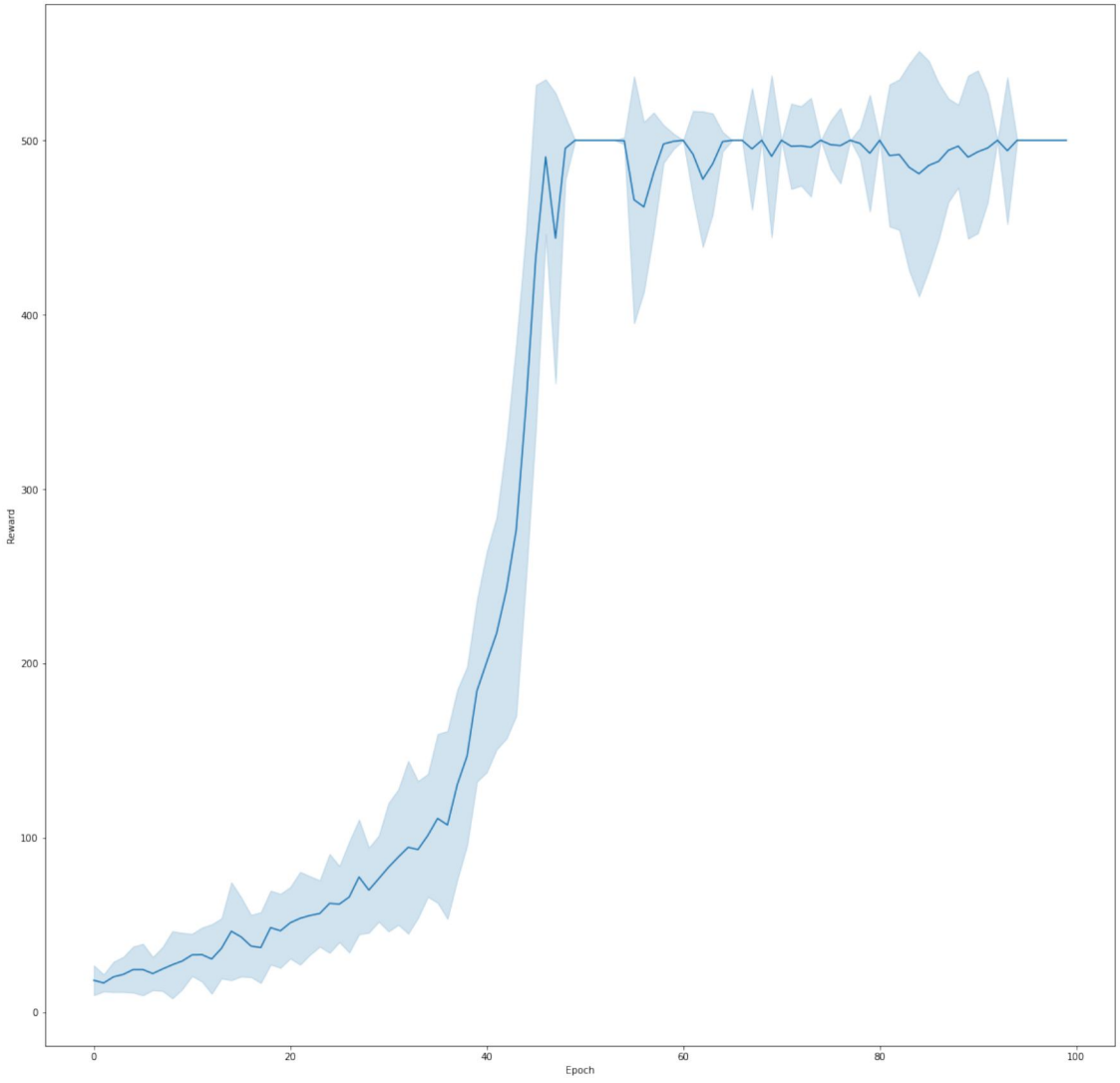
```
        loss.backward()
        self.policy_optimizer.step()
        return total_rewards
```

[ ]:
```
# You will be graded on this output this cell, so kindly run it
# keep the same config
config = {
    'env_id': 'CartPole-v1',
    'seed': 8953,
    'gamma': 1.0,
    'policy_layers': [16, 8],
    'policy_learning_rate': 1e-2,
    'use_baseline': False,
}
agent = REINFORCEv2Agent(config)
REINFORCEv2_rewards = agent.train(n_episodes=50, n_iterations=100)
```

```
the device is: cpu
Iteration 1/100: rewards 18.28 +/- 8.44
Iteration 2/100: rewards 16.74 +/- 4.84
Iteration 3/100: rewards 20.24 +/- 8.6
Iteration 4/100: rewards 21.66 +/- 9.97
Iteration 5/100: rewards 24.42 +/- 13.08
Iteration 6/100: rewards 24.42 +/- 14.64
Iteration 7/100: rewards 22.14 +/- 9.41
Iteration 8/100: rewards 24.78 +/- 12.52
Iteration 9/100: rewards 27.2 +/- 19.1
Iteration 10/100: rewards 29.32 +/- 15.98
Iteration 11/100: rewards 32.84 +/- 12.01
Iteration 12/100: rewards 33.02 +/- 15.31
Iteration 13/100: rewards 30.5 +/- 19.73
Iteration 14/100: rewards 36.62 +/- 17.06
Iteration 15/100: rewards 46.4 +/- 27.78
Iteration 16/100: rewards 43.14 +/- 22.45
Iteration 17/100: rewards 37.94 +/- 17.64
Iteration 18/100: rewards 37.04 +/- 20.1
Iteration 19/100: rewards 48.48 +/- 20.98
Iteration 20/100: rewards 46.66 +/- 20.99
Iteration 21/100: rewards 51.28 +/- 20.31
Iteration 22/100: rewards 53.82 +/- 26.35
Iteration 23/100: rewards 55.44 +/- 22.31
Iteration 24/100: rewards 56.58 +/- 18.84
Iteration 25/100: rewards 62.38 +/- 28.04
Iteration 26/100: rewards 61.98 +/- 21.61
Iteration 27/100: rewards 66.02 +/- 31.46
Iteration 28/100: rewards 77.52 +/- 32.58
Iteration 29/100: rewards 70.0 +/- 24.12
Iteration 30/100: rewards 76.52 +/- 24.43
```

```
Iteration 31/100: rewards 83.1 +/- 36.45
Iteration 32/100: rewards 88.96 +/- 38.48
Iteration 33/100: rewards 94.54 +/- 49.02
Iteration 34/100: rewards 93.26 +/- 38.87
Iteration 35/100: rewards 101.32 +/- 34.88
Iteration 36/100: rewards 111.1 +/- 47.92
Iteration 37/100: rewards 107.36 +/- 53.36
Iteration 38/100: rewards 130.48 +/- 53.93
Iteration 39/100: rewards 146.9 +/- 50.69
Iteration 40/100: rewards 184.28 +/- 51.58
Iteration 41/100: rewards 200.82 +/- 62.61
Iteration 42/100: rewards 217.22 +/- 65.9
Iteration 43/100: rewards 242.06 +/- 84.36
Iteration 44/100: rewards 276.5 +/- 105.55
Iteration 45/100: rewards 348.26 +/- 98.69
Iteration 46/100: rewards 433.26 +/- 97.44
Iteration 47/100: rewards 490.48 +/- 43.98
Iteration 48/100: rewards 443.86 +/- 82.43
Iteration 49/100: rewards 495.42 +/- 17.99
Iteration 50/100: rewards 500.0 +/- 0.0
Iteration 51/100: rewards 500.0 +/- 0.0
Iteration 52/100: rewards 500.0 +/- 0.0
Iteration 53/100: rewards 500.0 +/- 0.0
Iteration 54/100: rewards 500.0 +/- 0.0
Iteration 55/100: rewards 499.72 +/- 1.96
Iteration 56/100: rewards 465.86 +/- 70.12
Iteration 57/100: rewards 461.8 +/- 48.18
Iteration 58/100: rewards 481.48 +/- 34.09
Iteration 59/100: rewards 497.88 +/- 10.71
Iteration 60/100: rewards 499.36 +/- 4.48
Iteration 61/100: rewards 500.0 +/- 0.0
Iteration 62/100: rewards 492.0 +/- 24.64
Iteration 63/100: rewards 477.7 +/- 38.48
Iteration 64/100: rewards 486.48 +/- 28.63
Iteration 65/100: rewards 499.2 +/- 5.6
Iteration 66/100: rewards 500.0 +/- 0.0
Iteration 67/100: rewards 500.0 +/- 0.0
Iteration 68/100: rewards 495.06 +/- 34.58
Iteration 69/100: rewards 500.0 +/- 0.0
Iteration 70/100: rewards 490.78 +/- 46.0
Iteration 71/100: rewards 500.0 +/- 0.0
Iteration 72/100: rewards 496.54 +/- 24.22
Iteration 73/100: rewards 496.78 +/- 22.54
Iteration 74/100: rewards 496.0 +/- 28.0
Iteration 75/100: rewards 500.0 +/- 0.0
Iteration 76/100: rewards 497.48 +/- 13.56
Iteration 77/100: rewards 496.94 +/- 21.42
Iteration 78/100: rewards 500.0 +/- 0.0
```

```
Iteration 79/100: rewards 498.2 +/- 8.82
Iteration 80/100: rewards 492.56 +/- 33.17
Iteration 81/100: rewards 500.0 +/- 0.0
Iteration 82/100: rewards 491.28 +/- 40.32
Iteration 83/100: rewards 491.78 +/- 42.73
Iteration 84/100: rewards 484.52 +/- 58.73
Iteration 85/100: rewards 480.82 +/- 69.63
Iteration 86/100: rewards 485.58 +/- 59.36
Iteration 87/100: rewards 487.92 +/- 44.53
Iteration 88/100: rewards 494.26 +/- 29.44
Iteration 89/100: rewards 496.64 +/- 23.52
Iteration 90/100: rewards 490.32 +/- 46.25
Iteration 91/100: rewards 493.4 +/- 46.2
Iteration 92/100: rewards 495.58 +/- 30.94
Iteration 93/100: rewards 500.0 +/- 0.0
Iteration 94/100: rewards 494.04 +/- 41.72
Iteration 95/100: rewards 500.0 +/- 0.0
Iteration 96/100: rewards 500.0 +/- 0.0
Iteration 97/100: rewards 500.0 +/- 0.0
Iteration 98/100: rewards 500.0 +/- 0.0
Iteration 99/100: rewards 500.0 +/- 0.0
Iteration 100/100: rewards 500.0 +/- 0.0
```



[ ]:
```
# You will be graded on this output this cell, so kindly run it
agent.evaluate()
```

```
<IPython.core.display.HTML object>
```

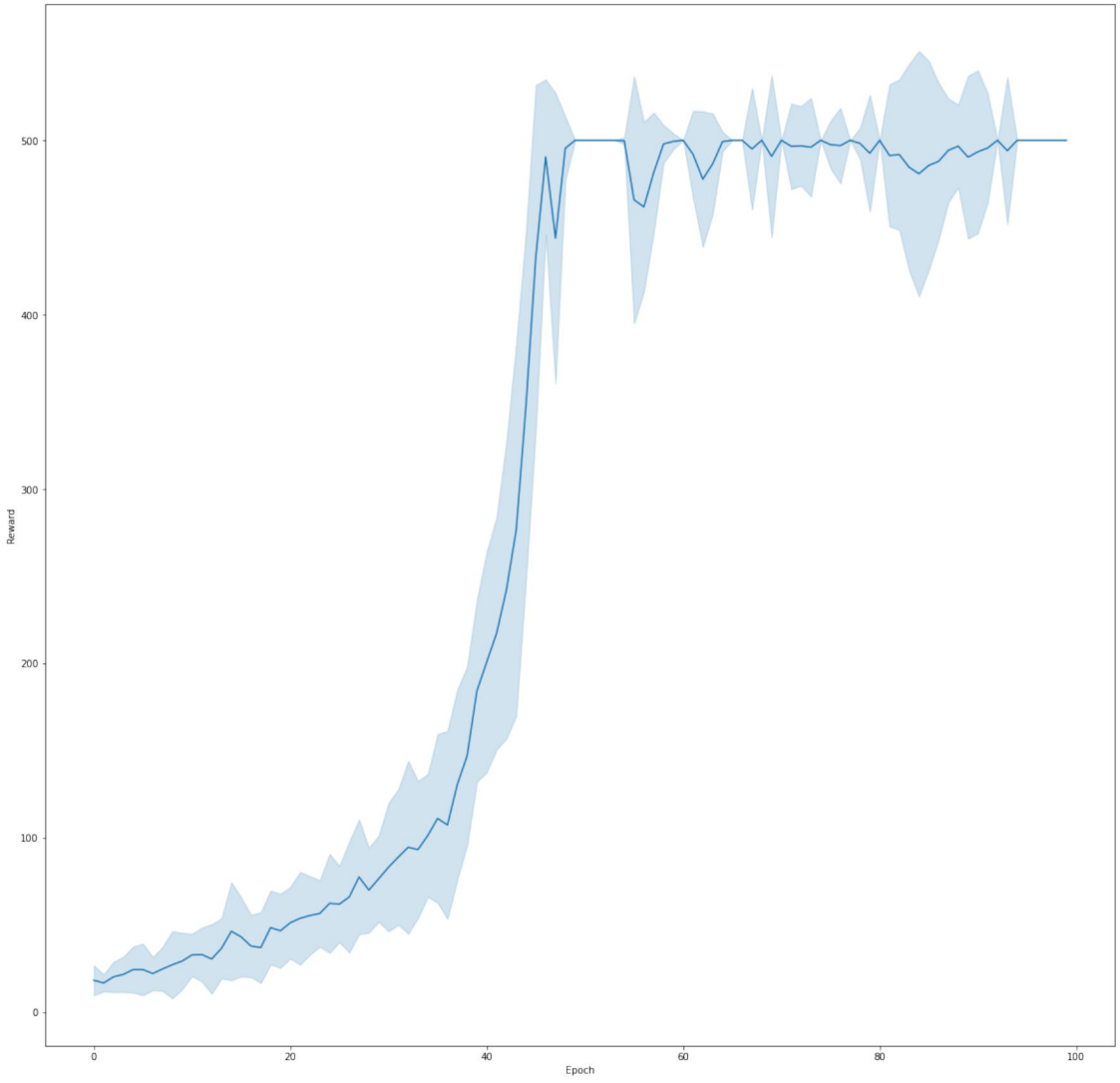
```
Reward: 500.0
```

4.2.2 Qn 1.2.b: Plot and compare the performance of the REINFORCEv1 and REINFORCEv2 agents for $\gamma = 1$. Report your observations and provide explanations for the same. [5 Marks]

[ ]:
```
# You will be graded on this output this cell, so kindly run it
fig, ax = plt.subplots()
BaseAgent.plot_rewards(REINFORCEv1_rewards, ax)
BaseAgent.plot_rewards(REINFORCEv2_rewards, ax)
```

2.1 Qn 1.2.a Implement REINFORCEv2 agent **10 / 10**

✓ **- 0 pts** **Correct**



[ ]:
```
# You will be graded on this output this cell, so kindly run it
agent.evaluate()
```

```
<IPython.core.display.HTML object>
```

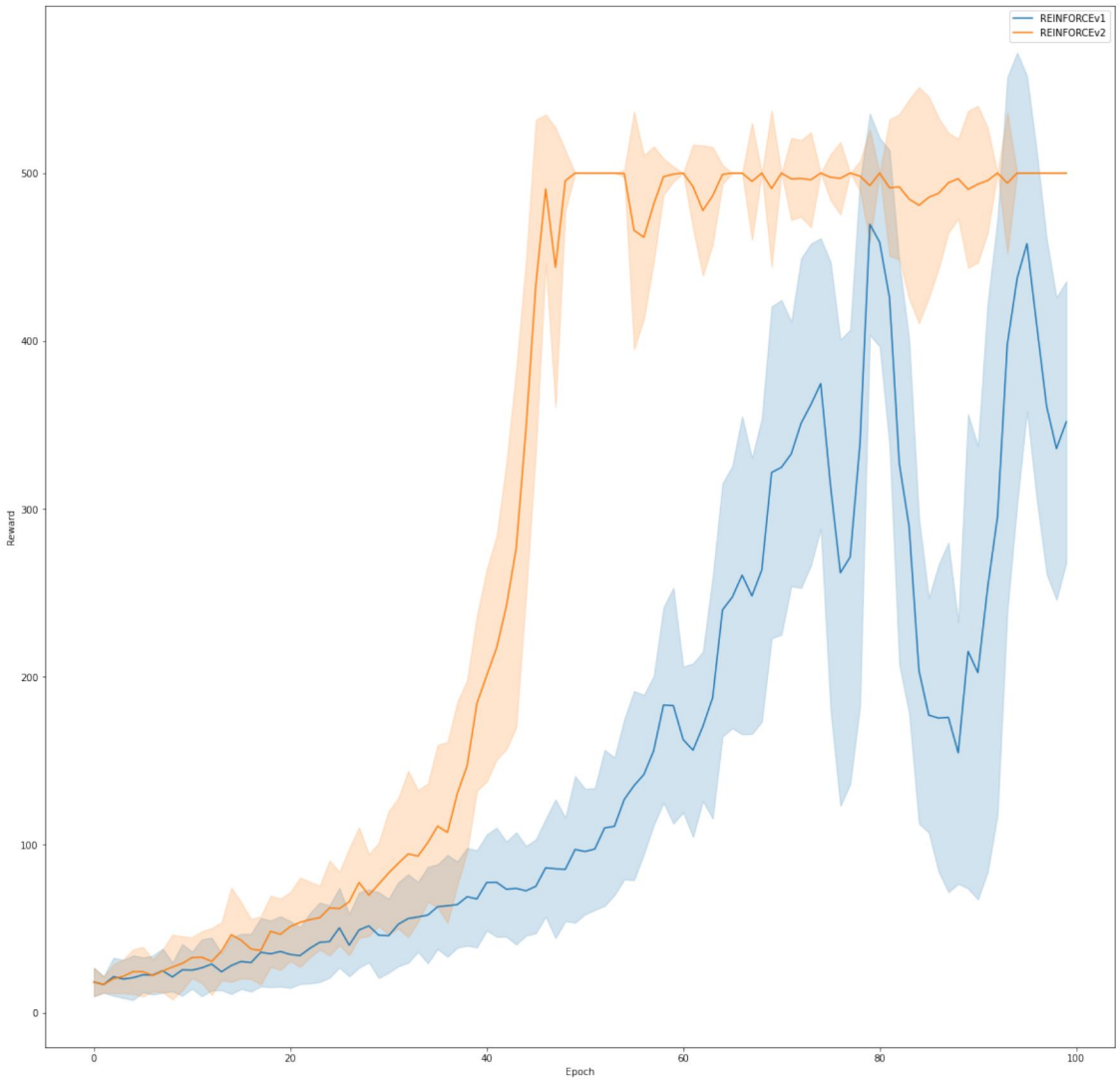
```
Reward: 500.0
```

4.2.2 Qn 1.2.b: Plot and compare the performance of the REINFORCEv1 and REINFORCEv2 agents for $\gamma = 1$. Report your observations and provide explanations for the same. [5 Marks]

[ ]:
```
# You will be graded on this output this cell, so kindly run it
fig, ax = plt.subplots()
BaseAgent.plot_rewards(REINFORCEv1_rewards, ax)
BaseAgent.plot_rewards(REINFORCEv2_rewards, ax)
```

```
plt.rcParams['figure.figsize'] = [20, 20]
plt.legend(labels=['REINFORCEv1', 'REINFORCEv2'])
```

[ ]: <matplotlib.legend.Legend at 0x7fec350b4e10>



We realize that Reinforce v2 converges faster than v1 and is more stable. We can also see that the variance is higher for v1 vs v2. This is due the factor Gt that is the only difference between the 2 algorithms. The difference in convergence is due to the fact that we have outlined previously. in v2 the learning rate is "adpative" and thus the algorithm, is more optimal in the case of v2 whereas in v1 the learning rate is constant. This difference leads to different results

2.2 Qn 1.2.b: Plot and compare the performance of the REINFORCEv1 and REINFORCEv2 agents for □=1 . **5 / 5**

✓ **- 0 pts** **Correct**

4.3 Qn 1.3 REINFORCE WITH baseline 25 Marks]

4.3.1 Qn 1.3.a Implement 'REINFORCEv2+B' agent as described below [15 Marks]

Implement a REINFORCE agent below with the following policy gradient computation.

$\nabla_\theta J(\theta) = \sum_j \sum_t (G_t^j - B(s_t^j)) \nabla_\theta ln \pi_\theta(a_t^j | s_t^j)$ \ where $G_t^j = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}^j$ is the discounted return computed starting from the current state, s_t^j for the episode j. \

Herein implement the baseline to be an estimator of the state-value function of the state at t, $B(s_t) = V(s_t)$. Towards that implement a value network with parameters, w to estimate the value of a state,i.e $B(s_t, w) = V(s_t)$.

Let's call this agent REINFORCEv2+B.

Note that you will be graded primarily on the output of the agent.train() and agent.evaluate() functions for this question.

[ ]:
```
# Insert your code and run this cell
class REINFORCEv2PlusBaselineAgent(BaseAgent):
    """ Baseline Agent:
        Here we try to reduce the variance by introducing a baseline, which is
        the value function in this case.
    """

    def optimize_model(self, n_episodes: int):
        """ YOU NEED TO IMPLEMENT THIS METHOD

            This method is called at each training iteration and is responsible␣
 ↪for
            (i) gathering a dataset of episodes
            (ii) computing the expectation of the policy gradient.
                 Note that you will only be computing the loss value

            In addition here, you will have to compute the loss of the value␣
 ↪function and
            call auto-diff on this loss to updae the parameters of the value␣
 ↪network.

            Here you have access to and need to make use of `self.value_model`
            and `self.value_optimizer`, and have to form a loss for updating the
            value function.

            HINT:
                * You need to use torch's `.detach()` to prevent re-flowing
                the gradients.
        """

        # ================================================================
```

```
        # INSERT YOUR CODE HERE !
        policy_loss = torch.tensor([0.0], requires_grad=True).to(self.device)
        value_loss = torch.tensor([0.0], requires_grad=True).to(self.device)
        total_rewards = np.empty(n_episodes)
        for episode in range(n_episodes):
          nested_loss = torch.tensor([0.0], requires_grad=True).to(self.device)
          states = []
          rewards = []
          actions = []
          sub_probs = []
          sub_values = []

          observation = self.monitor_env.reset()
          done = False

          while not done:
            states.append(observation)
            observation = torch.tensor(observation, dtype=torch.float)[None, :].
↪to(self.device)
            probs = self.policy_model.forward(observation)
            action = torch.multinomial(probs, 1)[0] # draw samples from dist
            sub_probs.append(torch.log(probs[0,int(action)]))
            value = self.value_model.forward(observation)
            sub_values.append(value)
            actions.append(action.detach().cpu().numpy())
            observation, reward, done, info = self.monitor_env.step(int(action))
            rewards.append(reward)


          total_rewards[episode] = sum(rewards)


          Gs = [r*self.gamma**i for i,r in enumerate(rewards)]
          Gk = np.cumsum(Gs[::-1])[::-1]

          for i in range(len(sub_probs)):

            policy_loss =  - (Gk[i]-sub_values[i])*sub_probs[i] + policy_loss
            value_loss = value_loss + (Gk[i]-sub_values[i])**2

        loss = policy_loss + value_loss
        self.monitor_env.close()

        # ============================================================================

        # self.policy_optimizer.zero_grad()
        # # policy_loss.backward()
```

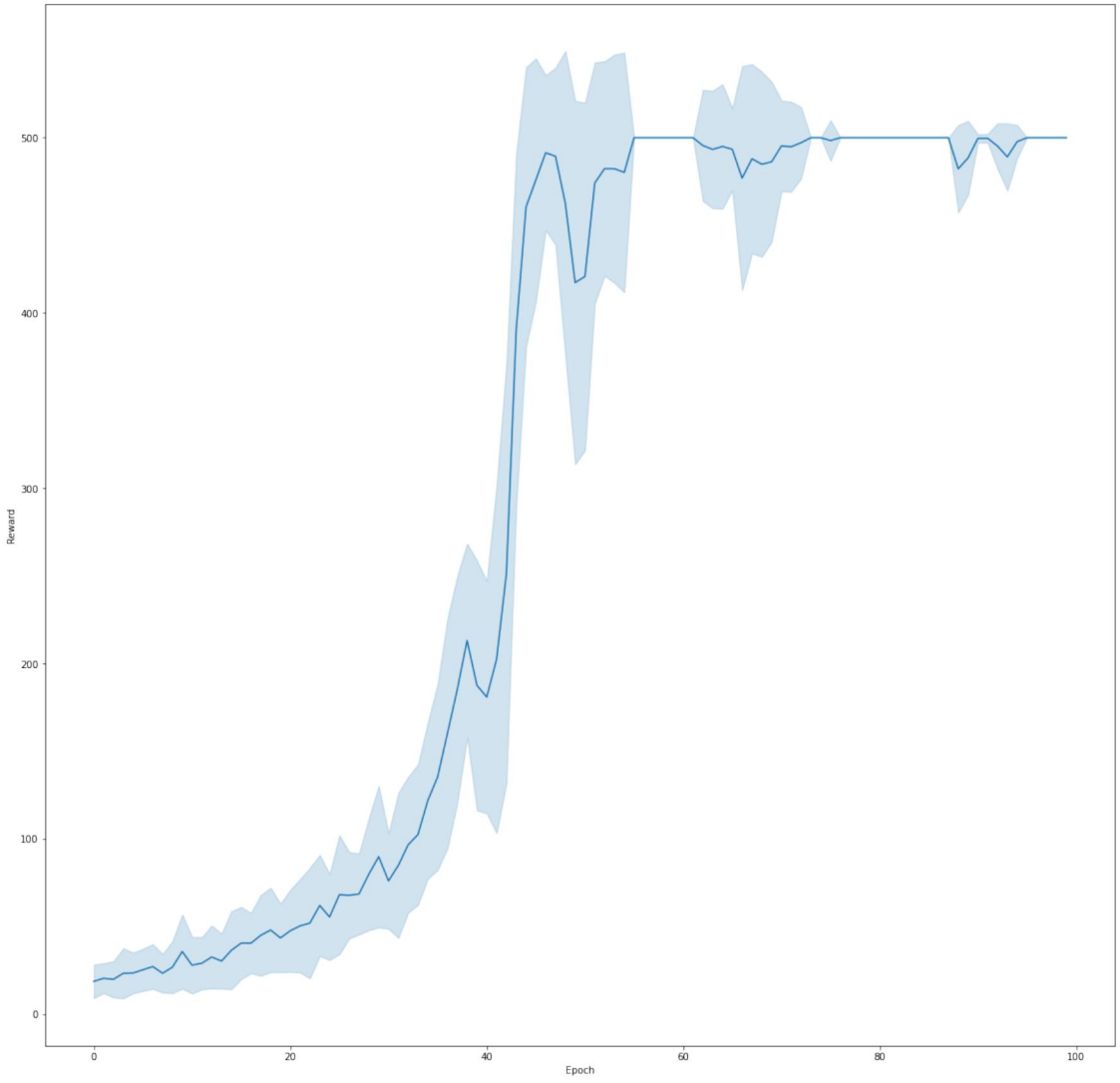
```
        # self.policy_optimizer.step()

        # # additionally we update the value network parameters
        # self.value_optimizer.zero_grad()
        # # value_loss.backward()
        # self.value_optimizer.step()
        self.policy_optimizer.zero_grad()
        self.value_optimizer.zero_grad()
        loss.backward()
        self.policy_optimizer.step()
        self.value_optimizer.step()


        return total_rewards
```

[ ]:
```
# You will be graded on this output this cell, so kindly run it.
# keep the config
config = {
    'env_id': 'CartPole-v1',
    'seed': 8953,
    'gamma': 1.0,
    'policy_layers': [16, 8],
    'policy_learning_rate': 1e-2,
    'use_baseline': True,
    'value_layers': [16, 8, 8],
    'value_learning_rate': 5e-3,
}
agent = REINFORCEv2PlusBaselineAgent(config)
REINFORCEv2PlusBaselineAgent_rewards = agent.train(n_episodes=50,␣
 ↪n_iterations=100)
```

```
the device is: cpu
Iteration 1/100: rewards 18.7 +/- 9.48
Iteration 2/100: rewards 20.44 +/- 8.46
Iteration 3/100: rewards 19.84 +/- 10.24
Iteration 4/100: rewards 23.28 +/- 14.17
Iteration 5/100: rewards 23.44 +/- 11.48
Iteration 6/100: rewards 25.28 +/- 11.94
Iteration 7/100: rewards 27.12 +/- 12.61
Iteration 8/100: rewards 23.3 +/- 10.88
Iteration 9/100: rewards 26.76 +/- 14.67
Iteration 10/100: rewards 35.7 +/- 20.98
Iteration 11/100: rewards 27.88 +/- 15.96
Iteration 12/100: rewards 29.02 +/- 14.78
Iteration 13/100: rewards 32.58 +/- 17.74
Iteration 14/100: rewards 30.24 +/- 15.51
Iteration 15/100: rewards 36.4 +/- 22.05
```

```
Iteration 16/100: rewards 40.5 +/- 20.39
Iteration 17/100: rewards 40.48 +/- 17.04
Iteration 18/100: rewards 44.96 +/- 22.86
Iteration 19/100: rewards 47.96 +/- 23.84
Iteration 20/100: rewards 43.44 +/- 19.4
Iteration 21/100: rewards 47.54 +/- 23.2
Iteration 22/100: rewards 50.34 +/- 26.43
Iteration 23/100: rewards 51.92 +/- 31.23
Iteration 24/100: rewards 61.98 +/- 28.57
Iteration 25/100: rewards 55.38 +/- 24.31
Iteration 26/100: rewards 68.1 +/- 33.62
Iteration 27/100: rewards 67.78 +/- 24.38
Iteration 28/100: rewards 68.54 +/- 22.9
Iteration 29/100: rewards 79.84 +/- 31.68
Iteration 30/100: rewards 89.8 +/- 39.96
Iteration 31/100: rewards 75.98 +/- 27.01
Iteration 32/100: rewards 84.82 +/- 41.0
Iteration 33/100: rewards 96.6 +/- 38.39
Iteration 34/100: rewards 102.42 +/- 39.64
Iteration 35/100: rewards 121.86 +/- 44.2
Iteration 36/100: rewards 135.28 +/- 52.57
Iteration 37/100: rewards 160.2 +/- 64.92
Iteration 38/100: rewards 185.2 +/- 64.1
Iteration 39/100: rewards 213.14 +/- 54.66
Iteration 40/100: rewards 187.54 +/- 70.56
Iteration 41/100: rewards 180.86 +/- 65.67
Iteration 42/100: rewards 202.4 +/- 98.16
Iteration 43/100: rewards 251.2 +/- 118.37
Iteration 44/100: rewards 390.2 +/- 99.57
Iteration 45/100: rewards 460.48 +/- 78.89
Iteration 46/100: rewards 475.88 +/- 68.6
Iteration 47/100: rewards 491.38 +/- 43.82
Iteration 48/100: rewards 489.38 +/- 50.01
Iteration 49/100: rewards 462.58 +/- 85.95
Iteration 50/100: rewards 417.3 +/- 102.65
Iteration 51/100: rewards 420.8 +/- 98.13
Iteration 52/100: rewards 474.16 +/- 68.05
Iteration 53/100: rewards 482.38 +/- 60.61
Iteration 54/100: rewards 482.3 +/- 64.47
Iteration 55/100: rewards 480.16 +/- 67.69
Iteration 56/100: rewards 500.0 +/- 0.0
Iteration 57/100: rewards 500.0 +/- 0.0
Iteration 58/100: rewards 500.0 +/- 0.0
Iteration 59/100: rewards 500.0 +/- 0.0
Iteration 60/100: rewards 500.0 +/- 0.0
Iteration 61/100: rewards 500.0 +/- 0.0
Iteration 62/100: rewards 500.0 +/- 0.0
Iteration 63/100: rewards 495.52 +/- 31.36
```

```
Iteration 64/100: rewards 493.28 +/- 33.19
Iteration 65/100: rewards 494.98 +/- 35.14
Iteration 66/100: rewards 493.36 +/- 23.23
Iteration 67/100: rewards 476.9 +/- 63.32
Iteration 68/100: rewards 487.96 +/- 53.46
Iteration 69/100: rewards 484.84 +/- 52.25
Iteration 70/100: rewards 486.22 +/- 45.25
Iteration 71/100: rewards 495.26 +/- 25.65
Iteration 72/100: rewards 494.82 +/- 25.39
Iteration 73/100: rewards 497.12 +/- 20.16
Iteration 74/100: rewards 500.0 +/- 0.0
Iteration 75/100: rewards 500.0 +/- 0.0
Iteration 76/100: rewards 498.36 +/- 11.48
Iteration 77/100: rewards 500.0 +/- 0.0
Iteration 78/100: rewards 500.0 +/- 0.0
Iteration 79/100: rewards 500.0 +/- 0.0
Iteration 80/100: rewards 500.0 +/- 0.0
Iteration 81/100: rewards 500.0 +/- 0.0
Iteration 82/100: rewards 500.0 +/- 0.0
Iteration 83/100: rewards 500.0 +/- 0.0
Iteration 84/100: rewards 500.0 +/- 0.0
Iteration 85/100: rewards 500.0 +/- 0.0
Iteration 86/100: rewards 500.0 +/- 0.0
Iteration 87/100: rewards 500.0 +/- 0.0
Iteration 88/100: rewards 500.0 +/- 0.0
Iteration 89/100: rewards 482.22 +/- 24.69
Iteration 90/100: rewards 488.44 +/- 20.91
Iteration 91/100: rewards 499.58 +/- 2.33
Iteration 92/100: rewards 499.64 +/- 2.52
Iteration 93/100: rewards 495.2 +/- 12.88
Iteration 94/100: rewards 489.04 +/- 18.99
Iteration 95/100: rewards 497.72 +/- 9.37
Iteration 96/100: rewards 500.0 +/- 0.0
Iteration 97/100: rewards 500.0 +/- 0.0
Iteration 98/100: rewards 500.0 +/- 0.0
Iteration 99/100: rewards 500.0 +/- 0.0
Iteration 100/100: rewards 500.0 +/- 0.0
```



[ ]:
```
# You will be graded on this output this cell, so kindly run this cell.
agent.evaluate()
```

```
<IPython.core.display.HTML object>
```

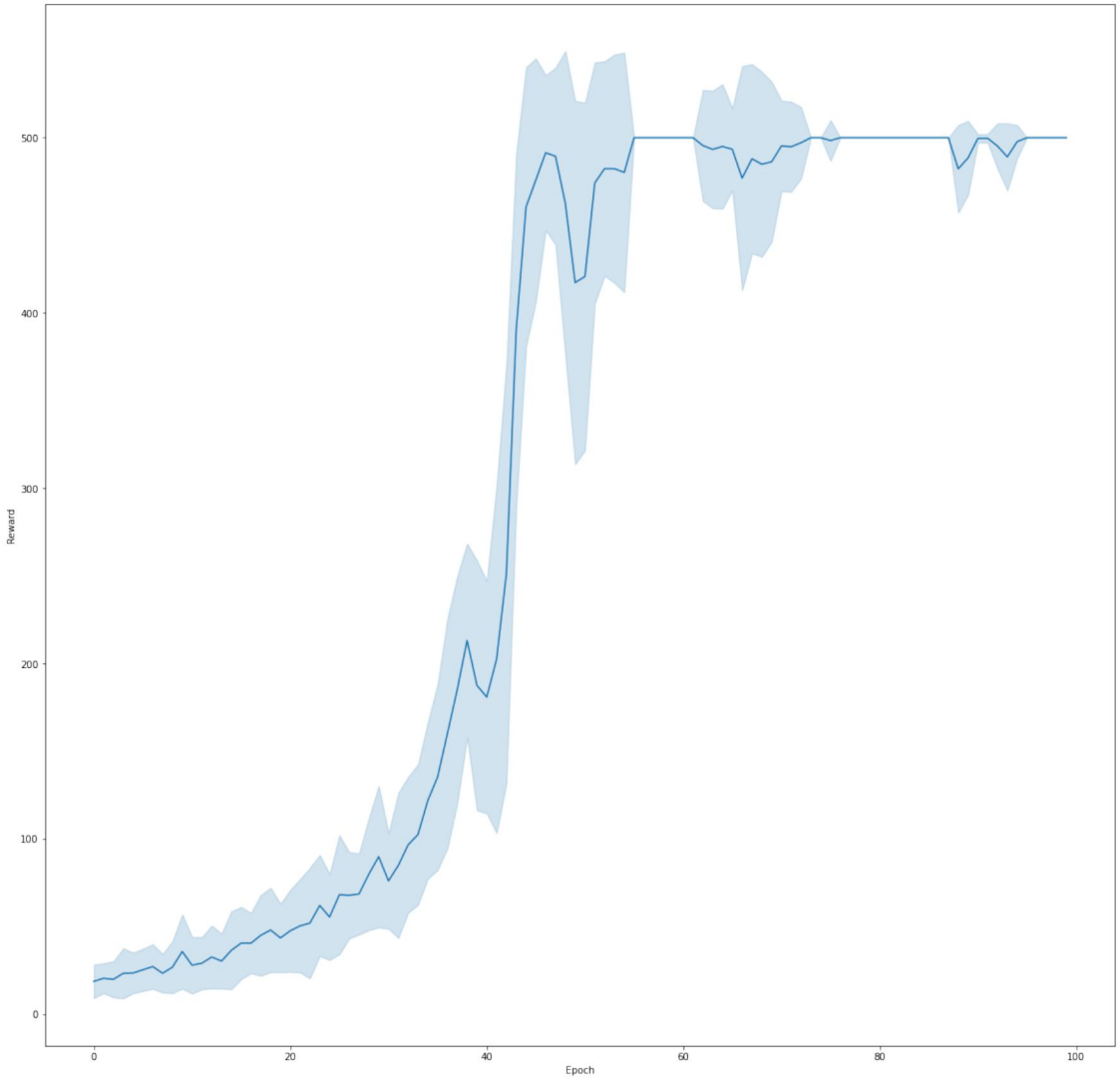
```
Reward: 500.0
```

4.3.2 Qn 1.3.b : Does introducing baselines have a meaning beyond variance reduction? [5 Marks]

Once again everything that has an effect on G_t has a consequence on the convergence of the algorithm. This is also the case of the baseline. adding a baseline here will improve convergence properties of the algorithm

3.1 Qn 1.3.a Implement 'REINFORCEv2+B' agent **15 / 15**

✓ **- 0 pts** **Correct**



[ ]:
```
# You will be graded on this output this cell, so kindly run this cell.
agent.evaluate()
```

```
<IPython.core.display.HTML object>
```

```
Reward: 500.0
```

4.3.2 Qn 1.3.b : Does introducing baselines have a meaning beyond variance reduction? [5 Marks]

Once again everything that has an effect on G_t has a consequence on the convergence of the algorithm. This is also the case of the baseline. adding a baseline here will improve convergence properties of the algorithm

3.2 Qn 1.3.b : Does introducing baselines have a meaning beyond variance reduction? 3 / 5

✓ - 0 pts Correct

✓ - 2 pts you should have mentioned something along the lines of the fact that the Q(s, a) - V(s) acts as an advantage estimator, A(s, a). Basically you need t mention that some states have higher state values by default and baseline is there to reduce this undesired effect.

4.3.3 Qn 1.3.c Plot and compare `REINFORCEv2+B` for $\gamma \in \{0.95, 0.975, 0.99, 0.995, 1\}$. [5 Marks]

Report your observations and explain the same.

[ ]:
```
# Insert your code here and run this cell
config = {
    'env_id': 'CartPole-v1',
    'seed': 8953,
    'gamma': 1.0,
    'policy_layers': [16, 8],
    'policy_learning_rate': 1e-2,
    'use_baseline': True,
    'value_layers': [16, 8, 8],
    'value_learning_rate': 5e-3,
}
gammas = [0.95,0.975,0.99,0.995,1]
dic = dict()
for gamma in gammas:
  config['gamma'] = gamma
  agent = REINFORCEv2PlusBaselineAgent(config)
  REINFORCEv2PlusBaselineAgent_rewards = agent.train(n_episodes=50,␣
 ↪n_iterations=100)
  dic[gamma] = REINFORCEv2PlusBaselineAgent_rewards

fig, ax = plt.subplots()
for k in list(dic.keys()):

  BaseAgent.plot_rewards(dic[k], ax)

plt.rcParams['figure.figsize'] = [20, 20]
plt.legend(labels=list(dic.keys()))
```

```
the device is: cpu
Iteration 1/100: rewards 18.7 +/- 9.48
Iteration 2/100: rewards 20.44 +/- 8.46
Iteration 3/100: rewards 19.6 +/- 8.78
Iteration 4/100: rewards 21.82 +/- 12.17
Iteration 5/100: rewards 24.58 +/- 12.02
Iteration 6/100: rewards 30.24 +/- 21.02
Iteration 7/100: rewards 26.96 +/- 18.63
Iteration 8/100: rewards 27.32 +/- 15.81
Iteration 9/100: rewards 34.54 +/- 21.52
Iteration 10/100: rewards 30.54 +/- 16.24
Iteration 11/100: rewards 29.58 +/- 14.42
Iteration 12/100: rewards 29.04 +/- 15.49
Iteration 13/100: rewards 29.98 +/- 14.56
Iteration 14/100: rewards 31.34 +/- 19.37
```

```
Iteration 15/100: rewards 36.76 +/- 21.62
Iteration 16/100: rewards 38.86 +/- 25.11
Iteration 17/100: rewards 38.7 +/- 21.35
Iteration 18/100: rewards 41.58 +/- 22.16
Iteration 19/100: rewards 42.44 +/- 21.07
Iteration 20/100: rewards 47.5 +/- 22.91
Iteration 21/100: rewards 44.34 +/- 23.43
Iteration 22/100: rewards 45.52 +/- 25.58
Iteration 23/100: rewards 52.26 +/- 25.56
Iteration 24/100: rewards 55.16 +/- 20.28
Iteration 25/100: rewards 52.02 +/- 24.03
Iteration 26/100: rewards 53.44 +/- 23.39
Iteration 27/100: rewards 58.06 +/- 23.18
Iteration 28/100: rewards 61.88 +/- 26.37
Iteration 29/100: rewards 60.26 +/- 22.62
Iteration 30/100: rewards 65.92 +/- 26.54
Iteration 31/100: rewards 61.88 +/- 25.95
Iteration 32/100: rewards 59.08 +/- 20.14
Iteration 33/100: rewards 74.04 +/- 37.47
Iteration 34/100: rewards 73.18 +/- 29.64
Iteration 35/100: rewards 69.84 +/- 24.43
Iteration 36/100: rewards 71.12 +/- 23.46
Iteration 37/100: rewards 80.16 +/- 36.33
Iteration 38/100: rewards 78.96 +/- 26.19
Iteration 39/100: rewards 88.1 +/- 48.39
Iteration 40/100: rewards 78.3 +/- 30.53
Iteration 41/100: rewards 75.58 +/- 23.58
Iteration 42/100: rewards 83.64 +/- 25.33
Iteration 43/100: rewards 75.68 +/- 26.49
Iteration 44/100: rewards 87.4 +/- 37.43
Iteration 45/100: rewards 80.54 +/- 25.19
Iteration 46/100: rewards 83.7 +/- 26.79
Iteration 47/100: rewards 92.04 +/- 29.89
Iteration 48/100: rewards 86.6 +/- 24.93
Iteration 49/100: rewards 88.38 +/- 31.08
Iteration 50/100: rewards 94.48 +/- 31.75
Iteration 51/100: rewards 93.9 +/- 30.04
Iteration 52/100: rewards 104.56 +/- 37.14
Iteration 53/100: rewards 100.96 +/- 31.21
Iteration 54/100: rewards 100.72 +/- 37.25
Iteration 55/100: rewards 107.12 +/- 30.85
Iteration 56/100: rewards 106.42 +/- 36.01
Iteration 57/100: rewards 103.94 +/- 42.9
Iteration 58/100: rewards 99.72 +/- 37.32
Iteration 59/100: rewards 118.98 +/- 44.75
Iteration 60/100: rewards 115.3 +/- 55.68
Iteration 61/100: rewards 124.54 +/- 40.58
Iteration 62/100: rewards 134.02 +/- 37.59
```

```
Iteration 63/100: rewards 141.34 +/- 55.4
Iteration 64/100: rewards 139.34 +/- 57.31
Iteration 65/100: rewards 111.2 +/- 46.79
Iteration 66/100: rewards 103.74 +/- 38.05
Iteration 67/100: rewards 115.18 +/- 61.31
Iteration 68/100: rewards 99.2 +/- 59.98
Iteration 69/100: rewards 86.92 +/- 34.39
Iteration 70/100: rewards 82.74 +/- 34.28
Iteration 71/100: rewards 83.16 +/- 38.05
Iteration 72/100: rewards 80.06 +/- 36.39
Iteration 73/100: rewards 95.36 +/- 54.29
Iteration 74/100: rewards 99.36 +/- 50.05
Iteration 75/100: rewards 101.92 +/- 59.01
Iteration 76/100: rewards 102.3 +/- 39.44
Iteration 77/100: rewards 115.7 +/- 46.71
Iteration 78/100: rewards 114.24 +/- 39.9
Iteration 79/100: rewards 118.16 +/- 47.78
Iteration 80/100: rewards 142.38 +/- 56.74
Iteration 81/100: rewards 142.1 +/- 48.58
Iteration 82/100: rewards 127.62 +/- 40.56
Iteration 83/100: rewards 119.48 +/- 49.01
Iteration 84/100: rewards 107.66 +/- 25.98
Iteration 85/100: rewards 107.24 +/- 36.15
Iteration 86/100: rewards 113.88 +/- 44.48
Iteration 87/100: rewards 113.6 +/- 39.99
Iteration 88/100: rewards 123.04 +/- 36.95
Iteration 89/100: rewards 121.72 +/- 37.0
Iteration 90/100: rewards 131.88 +/- 47.57
Iteration 91/100: rewards 140.38 +/- 65.74
Iteration 92/100: rewards 150.1 +/- 57.7
Iteration 93/100: rewards 142.06 +/- 39.25
Iteration 94/100: rewards 147.08 +/- 41.5
Iteration 95/100: rewards 137.56 +/- 36.62
Iteration 96/100: rewards 103.9 +/- 32.5
Iteration 97/100: rewards 98.16 +/- 25.19
Iteration 98/100: rewards 92.86 +/- 36.67
Iteration 99/100: rewards 92.98 +/- 29.11
Iteration 100/100: rewards 86.22 +/- 24.05
the device is: cpu
Iteration 1/100: rewards 18.7 +/- 9.48
Iteration 2/100: rewards 20.44 +/- 8.46
Iteration 3/100: rewards 19.84 +/- 9.73
Iteration 4/100: rewards 22.78 +/- 13.13
Iteration 5/100: rewards 23.94 +/- 10.49
Iteration 6/100: rewards 26.74 +/- 16.61
Iteration 7/100: rewards 22.44 +/- 10.86
Iteration 8/100: rewards 26.42 +/- 16.18
Iteration 9/100: rewards 26.86 +/- 15.21
```

```
Iteration 10/100: rewards 28.82 +/- 14.64
Iteration 11/100: rewards 32.22 +/- 18.21
Iteration 12/100: rewards 35.54 +/- 19.67
Iteration 13/100: rewards 30.8 +/- 14.34
Iteration 14/100: rewards 26.82 +/- 13.36
Iteration 15/100: rewards 41.26 +/- 24.82
Iteration 16/100: rewards 39.62 +/- 24.12
Iteration 17/100: rewards 40.14 +/- 25.22
Iteration 18/100: rewards 41.9 +/- 21.69
Iteration 19/100: rewards 48.44 +/- 24.48
Iteration 20/100: rewards 48.12 +/- 20.73
Iteration 21/100: rewards 52.04 +/- 23.54
Iteration 22/100: rewards 53.32 +/- 25.4
Iteration 23/100: rewards 51.2 +/- 24.1
Iteration 24/100: rewards 55.66 +/- 24.03
Iteration 25/100: rewards 57.96 +/- 20.82
Iteration 26/100: rewards 63.42 +/- 29.45
Iteration 27/100: rewards 64.36 +/- 25.47
Iteration 28/100: rewards 67.46 +/- 31.51
Iteration 29/100: rewards 64.98 +/- 25.09
Iteration 30/100: rewards 73.06 +/- 23.53
Iteration 31/100: rewards 69.88 +/- 27.71
Iteration 32/100: rewards 84.68 +/- 35.6
Iteration 33/100: rewards 79.84 +/- 32.21
Iteration 34/100: rewards 85.08 +/- 31.98
Iteration 35/100: rewards 85.62 +/- 33.61
Iteration 36/100: rewards 88.5 +/- 29.81
Iteration 37/100: rewards 86.46 +/- 36.28
Iteration 38/100: rewards 97.28 +/- 35.07
Iteration 39/100: rewards 107.24 +/- 31.15
Iteration 40/100: rewards 110.1 +/- 38.93
Iteration 41/100: rewards 110.02 +/- 33.91
Iteration 42/100: rewards 124.12 +/- 41.08
Iteration 43/100: rewards 132.08 +/- 47.81
Iteration 44/100: rewards 137.46 +/- 42.54
Iteration 45/100: rewards 163.9 +/- 66.83
Iteration 46/100: rewards 177.08 +/- 48.72
Iteration 47/100: rewards 175.56 +/- 56.48
Iteration 48/100: rewards 171.58 +/- 55.55
Iteration 49/100: rewards 162.78 +/- 63.81
Iteration 50/100: rewards 189.2 +/- 89.16
Iteration 51/100: rewards 215.32 +/- 79.95
Iteration 52/100: rewards 231.62 +/- 89.48
Iteration 53/100: rewards 259.46 +/- 79.31
Iteration 54/100: rewards 275.9 +/- 104.21
Iteration 55/100: rewards 305.88 +/- 91.87
Iteration 56/100: rewards 310.64 +/- 113.06
Iteration 57/100: rewards 298.6 +/- 89.68
```

```
Iteration 58/100: rewards 309.3 +/- 100.2
Iteration 59/100: rewards 360.72 +/- 110.78
Iteration 60/100: rewards 308.98 +/- 104.25
Iteration 61/100: rewards 327.12 +/- 79.42
Iteration 62/100: rewards 346.66 +/- 92.29
Iteration 63/100: rewards 339.02 +/- 91.71
Iteration 64/100: rewards 373.08 +/- 95.53
Iteration 65/100: rewards 331.8 +/- 83.99
Iteration 66/100: rewards 347.54 +/- 84.9
Iteration 67/100: rewards 286.4 +/- 117.0
Iteration 68/100: rewards 278.98 +/- 120.06
Iteration 69/100: rewards 246.3 +/- 111.97
Iteration 70/100: rewards 258.24 +/- 116.67
Iteration 71/100: rewards 248.42 +/- 105.43
Iteration 72/100: rewards 284.9 +/- 86.91
Iteration 73/100: rewards 266.88 +/- 87.91
Iteration 74/100: rewards 254.68 +/- 90.36
Iteration 75/100: rewards 300.74 +/- 86.97
Iteration 76/100: rewards 305.34 +/- 104.92
Iteration 77/100: rewards 328.84 +/- 101.96
Iteration 78/100: rewards 336.0 +/- 148.32
Iteration 79/100: rewards 296.82 +/- 143.52
Iteration 80/100: rewards 243.02 +/- 129.24
Iteration 81/100: rewards 389.74 +/- 141.81
Iteration 82/100: rewards 437.38 +/- 119.1
Iteration 83/100: rewards 420.78 +/- 134.91
Iteration 84/100: rewards 430.78 +/- 122.76
Iteration 85/100: rewards 457.68 +/- 92.24
Iteration 86/100: rewards 457.28 +/- 80.04
Iteration 87/100: rewards 443.4 +/- 86.13
Iteration 88/100: rewards 374.3 +/- 114.36
Iteration 89/100: rewards 230.42 +/- 78.97
Iteration 90/100: rewards 191.32 +/- 49.9
Iteration 91/100: rewards 155.34 +/- 28.33
Iteration 92/100: rewards 135.38 +/- 24.25
Iteration 93/100: rewards 120.16 +/- 17.94
Iteration 94/100: rewards 109.88 +/- 16.6
Iteration 95/100: rewards 96.22 +/- 16.47
Iteration 96/100: rewards 92.8 +/- 14.71
Iteration 97/100: rewards 88.24 +/- 13.45
Iteration 98/100: rewards 91.06 +/- 12.41
Iteration 99/100: rewards 91.26 +/- 15.2
Iteration 100/100: rewards 91.88 +/- 12.97
the device is: cpu
Iteration 1/100: rewards 18.7 +/- 9.48
Iteration 2/100: rewards 20.44 +/- 8.46
Iteration 3/100: rewards 21.26 +/- 11.34
Iteration 4/100: rewards 21.86 +/- 12.94
```

```
Iteration 5/100: rewards 23.26 +/- 9.96
Iteration 6/100: rewards 22.8 +/- 13.47
Iteration 7/100: rewards 26.58 +/- 12.92
Iteration 8/100: rewards 24.76 +/- 12.88
Iteration 9/100: rewards 33.76 +/- 18.86
Iteration 10/100: rewards 36.34 +/- 23.24
Iteration 11/100: rewards 31.92 +/- 25.3
Iteration 12/100: rewards 31.74 +/- 17.35
Iteration 13/100: rewards 30.74 +/- 20.76
Iteration 14/100: rewards 36.76 +/- 23.18
Iteration 15/100: rewards 39.06 +/- 19.93
Iteration 16/100: rewards 36.6 +/- 20.8
Iteration 17/100: rewards 38.5 +/- 21.41
Iteration 18/100: rewards 44.1 +/- 20.9
Iteration 19/100: rewards 37.96 +/- 15.79
Iteration 20/100: rewards 48.12 +/- 21.39
Iteration 21/100: rewards 49.74 +/- 24.48
Iteration 22/100: rewards 50.7 +/- 21.51
Iteration 23/100: rewards 54.96 +/- 27.33
Iteration 24/100: rewards 58.32 +/- 23.54
Iteration 25/100: rewards 62.96 +/- 35.16
Iteration 26/100: rewards 67.42 +/- 30.25
Iteration 27/100: rewards 68.26 +/- 26.36
Iteration 28/100: rewards 69.8 +/- 24.83
Iteration 29/100: rewards 80.18 +/- 34.55
Iteration 30/100: rewards 80.12 +/- 34.56
Iteration 31/100: rewards 79.8 +/- 47.94
Iteration 32/100: rewards 79.16 +/- 26.14
Iteration 33/100: rewards 83.06 +/- 25.42
Iteration 34/100: rewards 105.22 +/- 44.56
Iteration 35/100: rewards 106.7 +/- 45.22
Iteration 36/100: rewards 122.98 +/- 57.06
Iteration 37/100: rewards 125.84 +/- 51.35
Iteration 38/100: rewards 165.1 +/- 65.24
Iteration 39/100: rewards 185.32 +/- 66.61
Iteration 40/100: rewards 184.0 +/- 62.75
Iteration 41/100: rewards 215.66 +/- 76.83
Iteration 42/100: rewards 194.96 +/- 66.93
Iteration 43/100: rewards 217.06 +/- 81.38
Iteration 44/100: rewards 238.22 +/- 89.85
Iteration 45/100: rewards 287.76 +/- 89.15
Iteration 46/100: rewards 324.7 +/- 108.34
Iteration 47/100: rewards 321.4 +/- 101.76
Iteration 48/100: rewards 389.3 +/- 110.82
Iteration 49/100: rewards 445.9 +/- 93.43
Iteration 50/100: rewards 479.84 +/- 60.33
Iteration 51/100: rewards 477.68 +/- 78.43
Iteration 52/100: rewards 489.04 +/- 42.76
```

```
Iteration 53/100: rewards 475.64 +/- 72.05
Iteration 54/100: rewards 481.96 +/- 67.61
Iteration 55/100: rewards 480.98 +/- 52.99
Iteration 56/100: rewards 475.14 +/- 83.19
Iteration 57/100: rewards 489.78 +/- 43.73
Iteration 58/100: rewards 473.9 +/- 78.88
Iteration 59/100: rewards 498.66 +/- 9.38
Iteration 60/100: rewards 486.38 +/- 54.25
Iteration 61/100: rewards 488.12 +/- 62.43
Iteration 62/100: rewards 478.26 +/- 59.84
Iteration 63/100: rewards 498.18 +/- 12.74
Iteration 64/100: rewards 469.54 +/- 78.91
Iteration 65/100: rewards 453.08 +/- 101.39
Iteration 66/100: rewards 459.08 +/- 84.73
Iteration 67/100: rewards 454.26 +/- 90.41
Iteration 68/100: rewards 454.5 +/- 82.43
Iteration 69/100: rewards 420.5 +/- 116.38
Iteration 70/100: rewards 411.18 +/- 109.98
Iteration 71/100: rewards 421.78 +/- 111.44
Iteration 72/100: rewards 468.92 +/- 55.47
Iteration 73/100: rewards 409.9 +/- 110.05
Iteration 74/100: rewards 368.2 +/- 95.41
Iteration 75/100: rewards 365.4 +/- 104.59
Iteration 76/100: rewards 395.12 +/- 92.74
Iteration 77/100: rewards 410.28 +/- 88.28
Iteration 78/100: rewards 455.66 +/- 77.92
Iteration 79/100: rewards 487.36 +/- 49.2
Iteration 80/100: rewards 492.52 +/- 29.29
Iteration 81/100: rewards 477.94 +/- 76.97
Iteration 82/100: rewards 478.2 +/- 75.79
Iteration 83/100: rewards 495.52 +/- 29.69
Iteration 84/100: rewards 492.84 +/- 50.12
Iteration 85/100: rewards 490.14 +/- 49.43
Iteration 86/100: rewards 488.84 +/- 55.73
Iteration 87/100: rewards 500.0 +/- 0.0
Iteration 88/100: rewards 500.0 +/- 0.0
Iteration 89/100: rewards 497.38 +/- 18.34
Iteration 90/100: rewards 500.0 +/- 0.0
Iteration 91/100: rewards 491.5 +/- 41.68
Iteration 92/100: rewards 500.0 +/- 0.0
Iteration 93/100: rewards 500.0 +/- 0.0
Iteration 94/100: rewards 500.0 +/- 0.0
Iteration 95/100: rewards 500.0 +/- 0.0
Iteration 96/100: rewards 500.0 +/- 0.0
Iteration 97/100: rewards 500.0 +/- 0.0
Iteration 98/100: rewards 500.0 +/- 0.0
Iteration 99/100: rewards 500.0 +/- 0.0
Iteration 100/100: rewards 500.0 +/- 0.0
```

```
the device is: cpu
Iteration 1/100: rewards 18.7 +/- 9.48
Iteration 2/100: rewards 20.44 +/- 8.46
Iteration 3/100: rewards 20.8 +/- 10.73
Iteration 4/100: rewards 22.32 +/- 11.63
Iteration 5/100: rewards 23.44 +/- 11.48
Iteration 6/100: rewards 25.28 +/- 11.94
Iteration 7/100: rewards 28.18 +/- 15.23
Iteration 8/100: rewards 24.78 +/- 13.81
Iteration 9/100: rewards 31.82 +/- 16.5
Iteration 10/100: rewards 31.36 +/- 16.91
Iteration 11/100: rewards 32.18 +/- 15.6
Iteration 12/100: rewards 31.88 +/- 19.35
Iteration 13/100: rewards 32.58 +/- 20.94
Iteration 14/100: rewards 34.36 +/- 21.91
Iteration 15/100: rewards 41.5 +/- 18.25
Iteration 16/100: rewards 44.2 +/- 23.18
Iteration 17/100: rewards 43.64 +/- 22.28
Iteration 18/100: rewards 47.46 +/- 24.56
Iteration 19/100: rewards 43.88 +/- 21.94
Iteration 20/100: rewards 53.66 +/- 31.71
Iteration 21/100: rewards 55.14 +/- 25.87
Iteration 22/100: rewards 53.1 +/- 30.12
Iteration 23/100: rewards 54.34 +/- 22.45
Iteration 24/100: rewards 65.26 +/- 21.44
Iteration 25/100: rewards 67.88 +/- 28.58
Iteration 26/100: rewards 64.36 +/- 27.28
Iteration 27/100: rewards 71.12 +/- 25.0
Iteration 28/100: rewards 78.5 +/- 26.25
Iteration 29/100: rewards 84.4 +/- 30.57
Iteration 30/100: rewards 86.84 +/- 34.53
Iteration 31/100: rewards 94.64 +/- 36.56
Iteration 32/100: rewards 117.8 +/- 58.24
Iteration 33/100: rewards 123.28 +/- 57.02
Iteration 34/100: rewards 134.12 +/- 61.16
Iteration 35/100: rewards 172.22 +/- 68.06
Iteration 36/100: rewards 172.22 +/- 63.14
Iteration 37/100: rewards 203.4 +/- 70.27
Iteration 38/100: rewards 228.1 +/- 86.05
Iteration 39/100: rewards 250.08 +/- 131.5
Iteration 40/100: rewards 307.2 +/- 139.4
Iteration 41/100: rewards 405.14 +/- 124.07
Iteration 42/100: rewards 444.0 +/- 106.83
Iteration 43/100: rewards 411.22 +/- 112.03
Iteration 44/100: rewards 383.16 +/- 103.7
Iteration 45/100: rewards 428.26 +/- 81.14
Iteration 46/100: rewards 454.94 +/- 115.75
Iteration 47/100: rewards 491.04 +/- 57.8
```

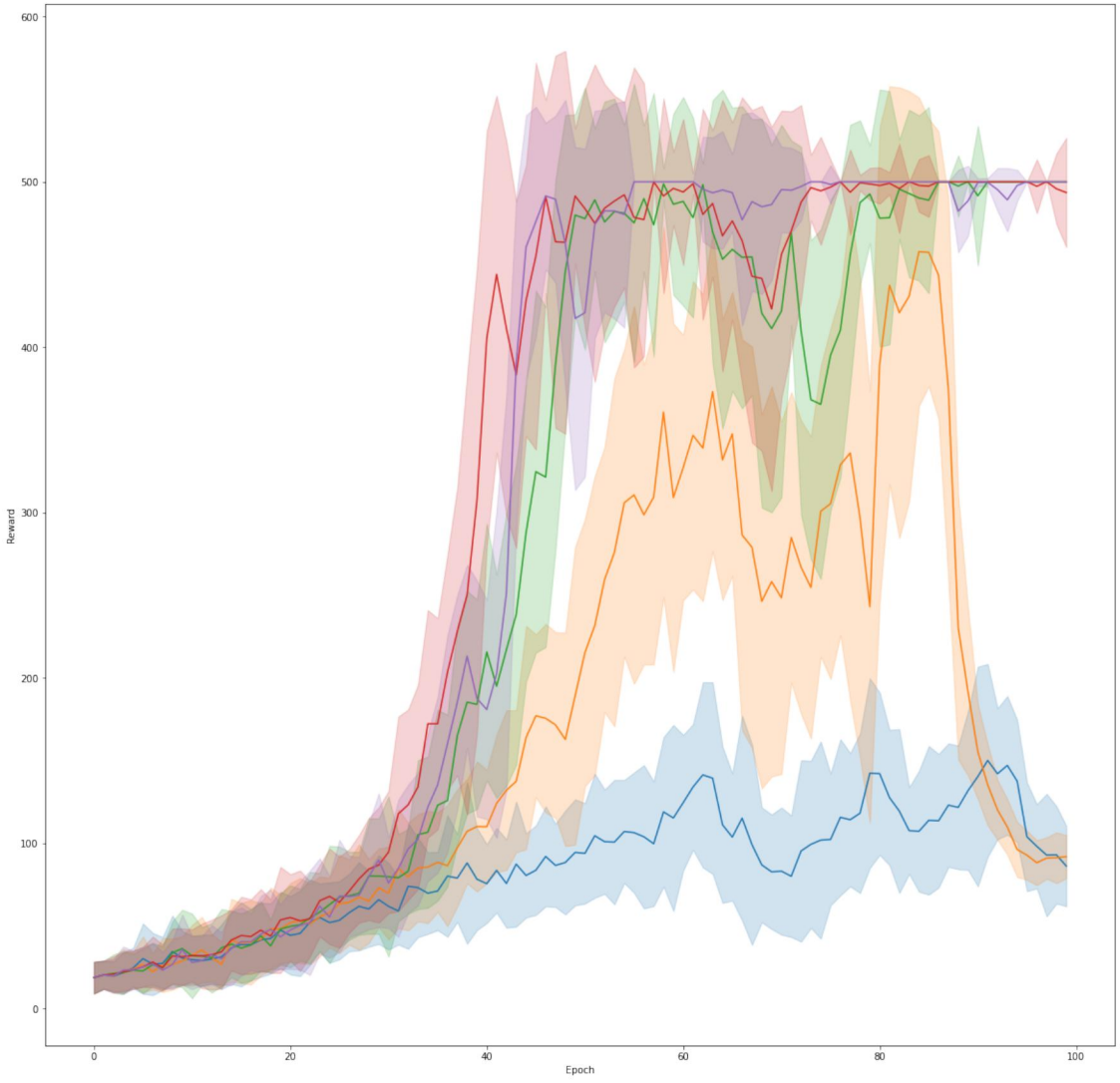
```
Iteration 48/100: rewards 463.66 +/- 111.41
Iteration 49/100: rewards 463.34 +/- 114.63
Iteration 50/100: rewards 491.34 +/- 40.26
Iteration 51/100: rewards 483.68 +/- 71.62
Iteration 52/100: rewards 474.86 +/- 95.03
Iteration 53/100: rewards 483.98 +/- 74.29
Iteration 54/100: rewards 488.3 +/- 63.06
Iteration 55/100: rewards 492.08 +/- 55.44
Iteration 56/100: rewards 478.4 +/- 89.78
Iteration 57/100: rewards 477.06 +/- 81.76
Iteration 58/100: rewards 500.0 +/- 0.0
Iteration 59/100: rewards 491.36 +/- 58.37
Iteration 60/100: rewards 495.96 +/- 21.91
Iteration 61/100: rewards 493.76 +/- 43.68
Iteration 62/100: rewards 498.74 +/- 6.23
Iteration 63/100: rewards 480.24 +/- 63.09
Iteration 64/100: rewards 486.84 +/- 43.94
Iteration 65/100: rewards 467.28 +/- 81.35
Iteration 66/100: rewards 476.3 +/- 58.97
Iteration 67/100: rewards 464.34 +/- 85.84
Iteration 68/100: rewards 442.8 +/- 99.49
Iteration 69/100: rewards 441.52 +/- 103.26
Iteration 70/100: rewards 423.02 +/- 109.11
Iteration 71/100: rewards 456.1 +/- 85.98
Iteration 72/100: rewards 470.24 +/- 71.56
Iteration 73/100: rewards 487.58 +/- 58.1
Iteration 74/100: rewards 496.4 +/- 19.71
Iteration 75/100: rewards 494.44 +/- 32.32
Iteration 76/100: rewards 496.72 +/- 16.51
Iteration 77/100: rewards 500.0 +/- 0.0
Iteration 78/100: rewards 493.64 +/- 25.54
Iteration 79/100: rewards 499.34 +/- 4.62
Iteration 80/100: rewards 498.62 +/- 9.66
Iteration 81/100: rewards 497.7 +/- 10.9
Iteration 82/100: rewards 499.06 +/- 6.58
Iteration 83/100: rewards 495.88 +/- 26.78
Iteration 84/100: rewards 500.0 +/- 0.0
Iteration 85/100: rewards 497.74 +/- 15.82
Iteration 86/100: rewards 497.36 +/- 18.48
Iteration 87/100: rewards 500.0 +/- 0.0
Iteration 88/100: rewards 500.0 +/- 0.0
Iteration 89/100: rewards 500.0 +/- 0.0
Iteration 90/100: rewards 500.0 +/- 0.0
Iteration 91/100: rewards 500.0 +/- 0.0
Iteration 92/100: rewards 500.0 +/- 0.0
Iteration 93/100: rewards 500.0 +/- 0.0
Iteration 94/100: rewards 500.0 +/- 0.0
Iteration 95/100: rewards 500.0 +/- 0.0
```

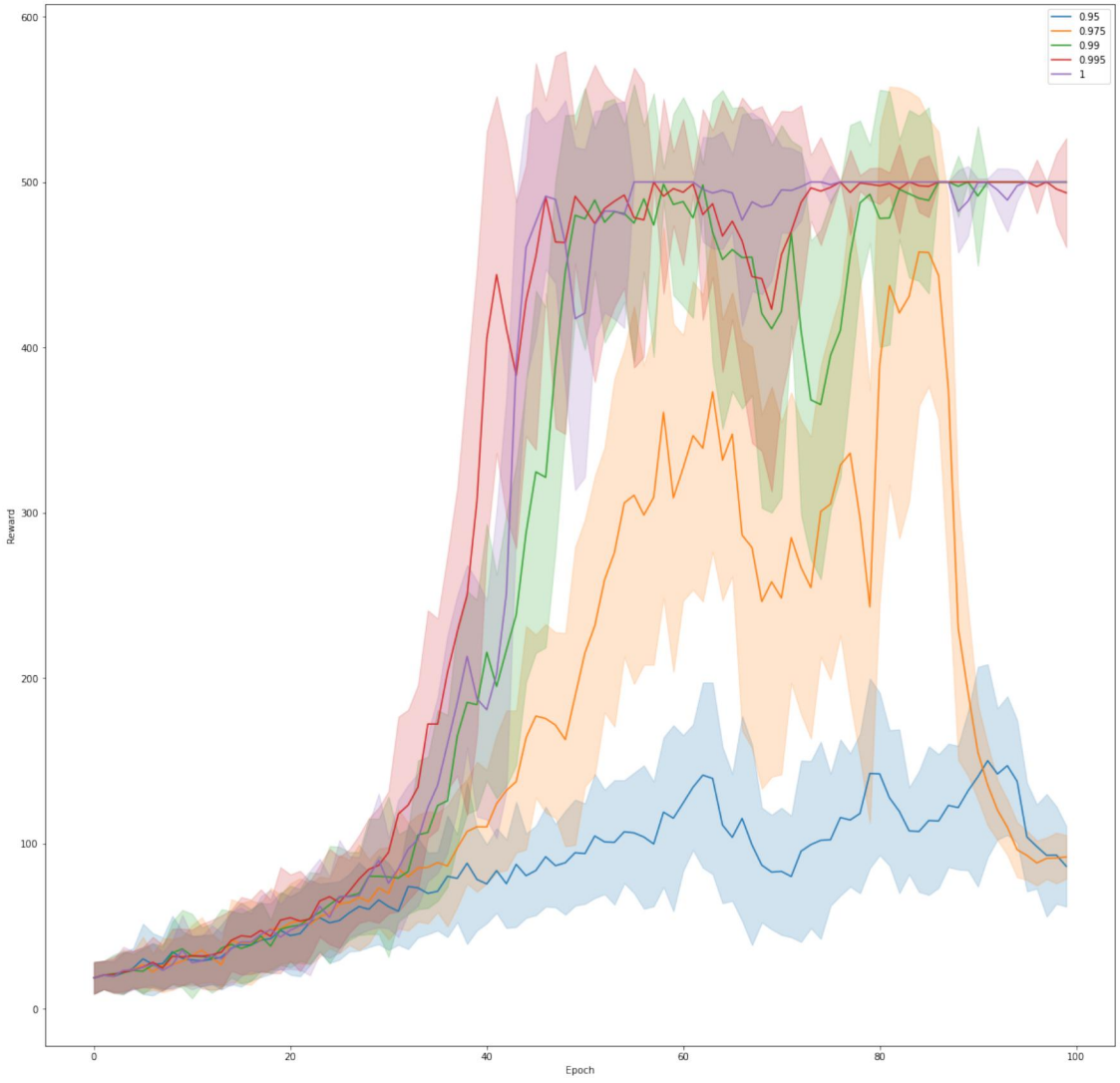
```
Iteration 96/100: rewards 500.0 +/- 0.0
Iteration 97/100: rewards 497.24 +/- 16.13
Iteration 98/100: rewards 500.0 +/- 0.0
Iteration 99/100: rewards 495.7 +/- 21.19
Iteration 100/100: rewards 493.44 +/- 32.69
the device is: cpu
Iteration 1/100: rewards 18.7 +/- 9.48
Iteration 2/100: rewards 20.44 +/- 8.46
Iteration 3/100: rewards 19.84 +/- 10.24
Iteration 4/100: rewards 23.28 +/- 14.17
Iteration 5/100: rewards 23.44 +/- 11.48
Iteration 6/100: rewards 25.28 +/- 11.94
Iteration 7/100: rewards 27.12 +/- 12.61
Iteration 8/100: rewards 23.3 +/- 10.88
Iteration 9/100: rewards 26.76 +/- 14.67
Iteration 10/100: rewards 35.7 +/- 20.98
Iteration 11/100: rewards 27.88 +/- 15.96
Iteration 12/100: rewards 29.02 +/- 14.78
Iteration 13/100: rewards 32.58 +/- 17.74
Iteration 14/100: rewards 30.24 +/- 15.51
Iteration 15/100: rewards 36.4 +/- 22.05
Iteration 16/100: rewards 40.5 +/- 20.39
Iteration 17/100: rewards 40.48 +/- 17.04
Iteration 18/100: rewards 44.96 +/- 22.86
Iteration 19/100: rewards 47.96 +/- 23.84
Iteration 20/100: rewards 43.44 +/- 19.4
Iteration 21/100: rewards 47.54 +/- 23.2
Iteration 22/100: rewards 50.34 +/- 26.43
Iteration 23/100: rewards 51.92 +/- 31.23
Iteration 24/100: rewards 61.98 +/- 28.57
Iteration 25/100: rewards 55.38 +/- 24.31
Iteration 26/100: rewards 68.1 +/- 33.62
Iteration 27/100: rewards 67.78 +/- 24.38
Iteration 28/100: rewards 68.54 +/- 22.9
Iteration 29/100: rewards 79.84 +/- 31.68
Iteration 30/100: rewards 89.8 +/- 39.96
Iteration 31/100: rewards 75.98 +/- 27.01
Iteration 32/100: rewards 84.82 +/- 41.0
Iteration 33/100: rewards 96.6 +/- 38.39
Iteration 34/100: rewards 102.42 +/- 39.64
Iteration 35/100: rewards 121.86 +/- 44.2
Iteration 36/100: rewards 135.28 +/- 52.57
Iteration 37/100: rewards 160.2 +/- 64.92
Iteration 38/100: rewards 185.2 +/- 64.1
Iteration 39/100: rewards 213.14 +/- 54.66
Iteration 40/100: rewards 187.54 +/- 70.56
Iteration 41/100: rewards 180.86 +/- 65.67
Iteration 42/100: rewards 202.4 +/- 98.16
```

```
Iteration 43/100: rewards 251.2 +/- 118.37
Iteration 44/100: rewards 390.2 +/- 99.57
Iteration 45/100: rewards 460.48 +/- 78.89
Iteration 46/100: rewards 475.88 +/- 68.6
Iteration 47/100: rewards 491.38 +/- 43.82
Iteration 48/100: rewards 489.38 +/- 50.01
Iteration 49/100: rewards 462.58 +/- 85.95
Iteration 50/100: rewards 417.3 +/- 102.65
Iteration 51/100: rewards 420.8 +/- 98.13
Iteration 52/100: rewards 474.16 +/- 68.05
Iteration 53/100: rewards 482.38 +/- 60.61
Iteration 54/100: rewards 482.3 +/- 64.47
Iteration 55/100: rewards 480.16 +/- 67.69
Iteration 56/100: rewards 500.0 +/- 0.0
Iteration 57/100: rewards 500.0 +/- 0.0
Iteration 58/100: rewards 500.0 +/- 0.0
Iteration 59/100: rewards 500.0 +/- 0.0
Iteration 60/100: rewards 500.0 +/- 0.0
Iteration 61/100: rewards 500.0 +/- 0.0
Iteration 62/100: rewards 500.0 +/- 0.0
Iteration 63/100: rewards 495.52 +/- 31.36
Iteration 64/100: rewards 493.28 +/- 33.19
Iteration 65/100: rewards 494.98 +/- 35.14
Iteration 66/100: rewards 493.36 +/- 23.23
Iteration 67/100: rewards 476.9 +/- 63.32
Iteration 68/100: rewards 487.96 +/- 53.46
Iteration 69/100: rewards 484.84 +/- 52.25
Iteration 70/100: rewards 486.22 +/- 45.25
Iteration 71/100: rewards 495.26 +/- 25.65
Iteration 72/100: rewards 494.82 +/- 25.39
Iteration 73/100: rewards 497.12 +/- 20.16
Iteration 74/100: rewards 500.0 +/- 0.0
Iteration 75/100: rewards 500.0 +/- 0.0
Iteration 76/100: rewards 498.36 +/- 11.48
Iteration 77/100: rewards 500.0 +/- 0.0
Iteration 78/100: rewards 500.0 +/- 0.0
Iteration 79/100: rewards 500.0 +/- 0.0
Iteration 80/100: rewards 500.0 +/- 0.0
Iteration 81/100: rewards 500.0 +/- 0.0
Iteration 82/100: rewards 500.0 +/- 0.0
Iteration 83/100: rewards 500.0 +/- 0.0
Iteration 84/100: rewards 500.0 +/- 0.0
Iteration 85/100: rewards 500.0 +/- 0.0
Iteration 86/100: rewards 500.0 +/- 0.0
Iteration 87/100: rewards 500.0 +/- 0.0
Iteration 88/100: rewards 500.0 +/- 0.0
Iteration 89/100: rewards 482.22 +/- 24.69
Iteration 90/100: rewards 488.44 +/- 20.91
```

```
Iteration 91/100: rewards 499.58 +/- 2.33
Iteration 92/100: rewards 499.64 +/- 2.52
Iteration 93/100: rewards 495.2 +/- 12.88
Iteration 94/100: rewards 489.04 +/- 18.99
Iteration 95/100: rewards 497.72 +/- 9.37
Iteration 96/100: rewards 500.0 +/- 0.0
Iteration 97/100: rewards 500.0 +/- 0.0
Iteration 98/100: rewards 500.0 +/- 0.0
Iteration 99/100: rewards 500.0 +/- 0.0
Iteration 100/100: rewards 500.0 +/- 0.0
```

[ ]: <matplotlib.legend.Legend at 0x7f63f83b4a90>





The effect of gamma is huge on the plots above. We can see that some algorithms converges really fast and very stable for the highest value of gamma while the results degrade with lower values of gamma. As we have seen previously, lower gamma leads a to a lower learning rate (indirectly) and we can see that on the figure above

5 Qn 2. ACTOR CRITIC [35 Marks]

5.0.1 Qn 2.1 Implement a one-step Actor-Critic agent below [15 Marks].

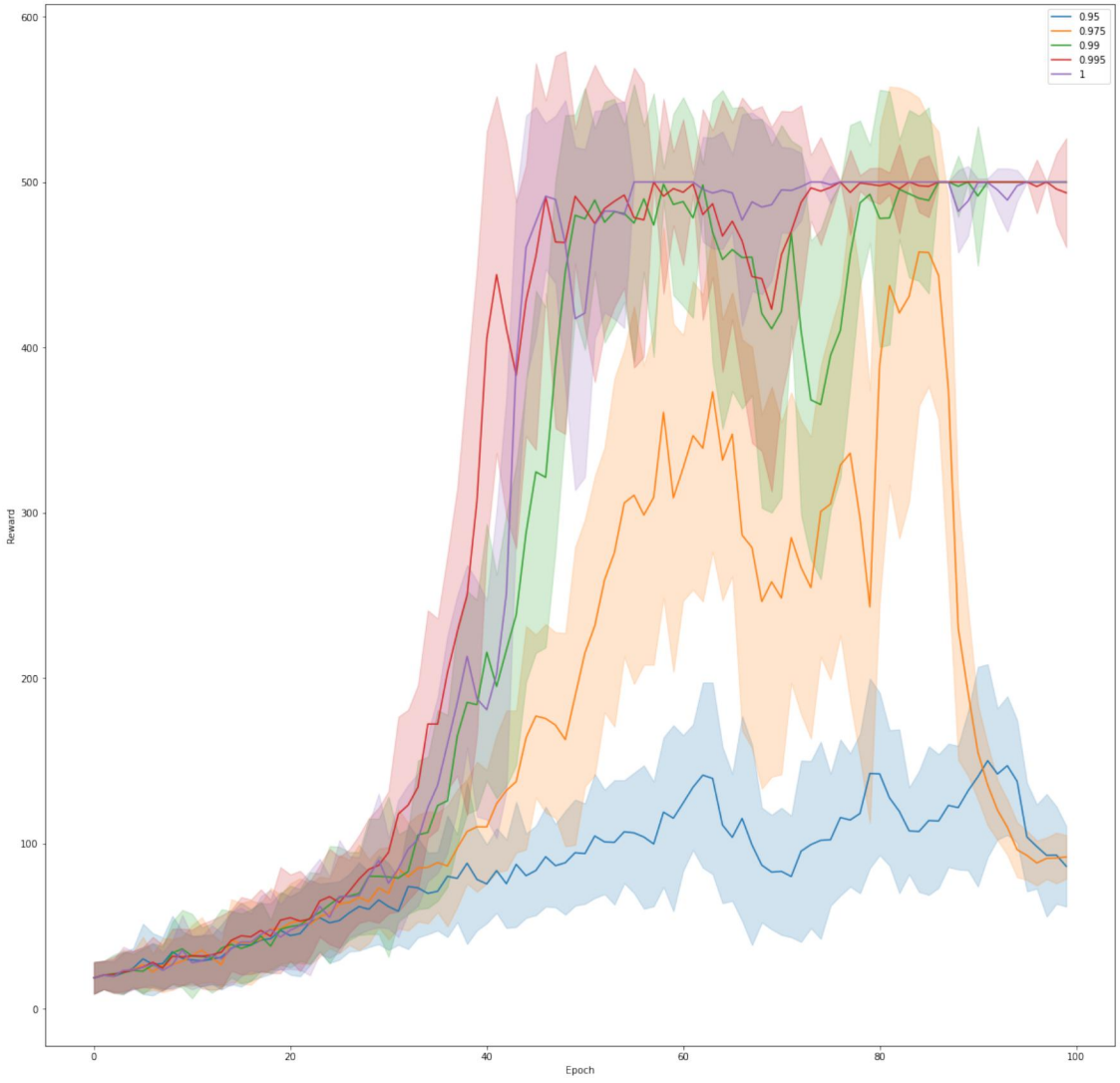
Implement an actor critic agent below with the following policy gradient computation.

$\nabla_\theta J(\theta) = \sum_j \sum_t (G^j_{t:t+1} - V(s^j_t)) \nabla_\theta ln \pi_\theta(a^j_t | s^j_t)$ \ where $G^j_{t:t+1} = R_t + \gamma V(s^j_{t+1})$ is the truncated one-step return computed starting from the current state, s^j_t for the episode j. \

Implement the critic network to be an estimator for state-value function.

3.3 Qn 1.3.c Plot and compare REINFORCEv2+B for $\gamma \in \{0.95, 0.975, 0.99, 0.995, 1\}$ **5 / 5**

✓ **- 0 pts Correct**



The effect of gamma is huge on the plots above. We can see that some algorithms converges really fast and very stable for the highest value of gamma while the results degrade with lower values of gamma. As we have seen previously, lower gamma leads a to a lower learning rate (indirectly) and we can see that on the figure above

5 Qn 2. ACTOR CRITIC [35 Marks]

5.0.1 Qn 2.1 Implement a one-step Actor-Critic agent below [15 Marks].

Implement an actor critic agent below with the following policy gradient computation.

$\nabla_\theta J(\theta) = \sum_j \sum_t (G^j_{t:t+1} - V(s^j_t)) \nabla_\theta ln \pi_\theta(a^j_t | s^j_t)$ \ where $G^j_{t:t+1} = R_t + \gamma V(s^j_{t+1})$ is the truncated one-step return computed starting from the current state, s^j_t for the episode j. \

Implement the critic network to be an estimator for state-value function.

Note that you will be graded primarily on the output of the agent.train() and agent.evaluate() functions for this question.

[ ]:
```
from itertools import count
# Insert your code and run this cell
class ActorCriticAgent(BaseAgent):
    """ A2C Agent: Actor-Critic
        Here we try to FURTHER reduce the variance via bootstrapping.
    """

    def optimize_model(self, n_episodes: int):
        """ YOU NEED TO IMPLEMENT THIS METHOD

            This method is called at each training iteration and is responsible␣
↪for
            (i) gathering a dataset of episodes
            (ii) computing the expectation of the policy gradient.
                 Note that you will only be computing the loss value
            In addition implement the critic network
            HINT:
                * If you've made it this far you don't need another hint!
        """
        # #␣
↪=====================================================================
        # #␣
↪=====================================================================
        # # INSERT YOUR CODE HERE !
        policy_loss = torch.tensor([0.0], requires_grad=True).to(self.device)
        value_loss = torch.tensor([0.0], requires_grad=True).to(self.device)
        total_rewards = np.empty(n_episodes)
        for episode in range(n_episodes):
          states = []
          rewards = []
          actions = []
          sub_probs = []
          sub_values = []

          observation = self.monitor_env.reset()
          done = False

          while not done:
            states.append(observation)
            observation = torch.tensor(observation, dtype=torch.float)[None, :].
↪to(self.device)
            value = self.value_model.forward(observation)
            probs = self.policy_model.forward(observation)
```

```
            action = torch.multinomial(probs, 1)[0] # draw samples from dist
            sub_probs.append(torch.log(probs[0,int(action)]))

            sub_values.append(value)
            actions.append(action.detach().cpu().numpy())
            observation, reward, done, info = self.monitor_env.step(int(action))
            rewards.append(reward)


          total_rewards[episode] = sum(rewards)


          # Gs = [r*self.gamma**i for i,r in enumerate(rewards)]
          # Gk = np.cumsum(Gs[::-1])[::-1]

          for i in range(len(sub_probs)):

            if i == (len(sub_probs)-1):
              one_step_return = rewards[i]
            else:
              one_step_return = rewards[i] + self.gamma *sub_values[i+1]

            advantage = (one_step_return-sub_values[i])
            policy_loss =  - advantage*sub_probs[i] + policy_loss

            value_loss = value_loss + (one_step_return - sub_values[i])**2

        loss = policy_loss + value_loss
        self.monitor_env.close()
        # ================================================================

        self.policy_optimizer.zero_grad()
        self.value_optimizer.zero_grad()
        # actor_loss.backward()
        loss.backward()
        # value_loss.backward()
        # critic_loss.backward()

        self.policy_optimizer.step()

        # self.value_optimizer.zero_grad()

        self.value_optimizer.step()
        return total_rewards
```

[ ]:
```
agent.monitor_env.close()
```

[ ]:
```
# new
config = {'env_id': 'CartPole-v1','seed': 8953,
          'gamma': 1,
          'policy_layers': [16, 8],
          'policy_learning_rate': 1e-3,
          'use_baseline': True,
          'value_layers': [16, 8],
          'value_learning_rate': 1e-2,}

agent = ActorCriticAgent(config)

ActorCritic_sum_rewards = agent.train(n_episodes=32, n_iterations=500)
```

```
the device is: cpu
Iteration 1/500: rewards 19.0 +/- 7.26
Iteration 2/500: rewards 18.59 +/- 9.64
Iteration 3/500: rewards 17.69 +/- 8.52
Iteration 4/500: rewards 20.94 +/- 6.98
Iteration 5/500: rewards 18.97 +/- 11.74
Iteration 6/500: rewards 18.75 +/- 9.46
Iteration 7/500: rewards 15.97 +/- 5.78
Iteration 8/500: rewards 18.84 +/- 7.36
Iteration 9/500: rewards 20.56 +/- 8.48
Iteration 10/500: rewards 18.84 +/- 6.64
Iteration 11/500: rewards 19.53 +/- 11.07
Iteration 12/500: rewards 20.56 +/- 9.37
Iteration 13/500: rewards 19.62 +/- 10.15
Iteration 14/500: rewards 17.38 +/- 7.72
Iteration 15/500: rewards 17.78 +/- 8.97
Iteration 16/500: rewards 21.03 +/- 12.03
Iteration 17/500: rewards 19.78 +/- 9.39
Iteration 18/500: rewards 18.44 +/- 8.89
Iteration 19/500: rewards 16.09 +/- 6.68
Iteration 20/500: rewards 19.28 +/- 8.8
Iteration 21/500: rewards 16.72 +/- 11.61
Iteration 22/500: rewards 18.47 +/- 8.72
Iteration 23/500: rewards 17.09 +/- 5.66
Iteration 24/500: rewards 18.97 +/- 8.46
Iteration 25/500: rewards 17.75 +/- 8.7
Iteration 26/500: rewards 18.19 +/- 10.91
Iteration 27/500: rewards 19.34 +/- 9.46
Iteration 28/500: rewards 21.78 +/- 9.82
Iteration 29/500: rewards 18.12 +/- 8.23
Iteration 30/500: rewards 22.22 +/- 14.37
Iteration 31/500: rewards 18.97 +/- 8.44
Iteration 32/500: rewards 19.03 +/- 12.09
Iteration 33/500: rewards 18.38 +/- 6.62
```

```
Iteration 34/500: rewards 18.78 +/- 9.11
Iteration 35/500: rewards 19.38 +/- 10.01
Iteration 36/500: rewards 20.62 +/- 10.23
Iteration 37/500: rewards 18.34 +/- 7.44
Iteration 38/500: rewards 21.75 +/- 10.44
Iteration 39/500: rewards 20.78 +/- 13.47
Iteration 40/500: rewards 19.34 +/- 11.23
Iteration 41/500: rewards 19.94 +/- 10.13
Iteration 42/500: rewards 24.38 +/- 14.8
Iteration 43/500: rewards 17.34 +/- 7.43
Iteration 44/500: rewards 19.78 +/- 8.97
Iteration 45/500: rewards 18.72 +/- 6.87
Iteration 46/500: rewards 21.19 +/- 10.28
Iteration 47/500: rewards 22.59 +/- 12.07
Iteration 48/500: rewards 19.22 +/- 9.08
Iteration 49/500: rewards 19.81 +/- 8.3
Iteration 50/500: rewards 21.41 +/- 13.3
Iteration 51/500: rewards 27.16 +/- 16.18
Iteration 52/500: rewards 22.97 +/- 12.29
Iteration 53/500: rewards 21.03 +/- 10.49
Iteration 54/500: rewards 21.03 +/- 9.75
Iteration 55/500: rewards 24.81 +/- 13.92
Iteration 56/500: rewards 22.88 +/- 11.24
Iteration 57/500: rewards 22.94 +/- 11.16
Iteration 58/500: rewards 19.25 +/- 9.32
Iteration 59/500: rewards 21.41 +/- 13.59
Iteration 60/500: rewards 25.72 +/- 14.93
Iteration 61/500: rewards 22.12 +/- 11.94
Iteration 62/500: rewards 24.62 +/- 10.77
Iteration 63/500: rewards 25.69 +/- 17.9
Iteration 64/500: rewards 19.31 +/- 6.32
Iteration 65/500: rewards 22.5 +/- 12.91
Iteration 66/500: rewards 22.03 +/- 10.24
Iteration 67/500: rewards 21.91 +/- 10.12
Iteration 68/500: rewards 25.12 +/- 14.05
Iteration 69/500: rewards 24.28 +/- 11.47
Iteration 70/500: rewards 25.97 +/- 14.91
Iteration 71/500: rewards 27.19 +/- 14.34
Iteration 72/500: rewards 25.91 +/- 15.56
Iteration 73/500: rewards 24.06 +/- 11.06
Iteration 74/500: rewards 30.31 +/- 15.69
Iteration 75/500: rewards 26.5 +/- 14.57
Iteration 76/500: rewards 25.31 +/- 13.02
Iteration 77/500: rewards 25.03 +/- 11.19
Iteration 78/500: rewards 26.31 +/- 14.5
Iteration 79/500: rewards 25.97 +/- 12.96
Iteration 80/500: rewards 27.97 +/- 19.85
Iteration 81/500: rewards 24.0 +/- 13.04
```

```
Iteration 82/500: rewards 24.72 +/- 14.74
Iteration 83/500: rewards 27.34 +/- 14.12
Iteration 84/500: rewards 22.16 +/- 11.56
Iteration 85/500: rewards 26.25 +/- 15.09
Iteration 86/500: rewards 27.72 +/- 18.89
Iteration 87/500: rewards 25.5 +/- 15.51
Iteration 88/500: rewards 23.94 +/- 9.06
Iteration 89/500: rewards 27.25 +/- 14.34
Iteration 90/500: rewards 29.69 +/- 12.34
Iteration 91/500: rewards 29.91 +/- 16.35
Iteration 92/500: rewards 29.12 +/- 20.09
Iteration 93/500: rewards 26.78 +/- 15.26
Iteration 94/500: rewards 25.78 +/- 10.94
Iteration 95/500: rewards 22.62 +/- 12.51
Iteration 96/500: rewards 28.56 +/- 11.64
Iteration 97/500: rewards 30.97 +/- 19.75
Iteration 98/500: rewards 25.25 +/- 12.68
Iteration 99/500: rewards 33.09 +/- 22.08
Iteration 100/500: rewards 27.09 +/- 16.2
Iteration 101/500: rewards 32.12 +/- 16.99
Iteration 102/500: rewards 27.66 +/- 11.79
Iteration 103/500: rewards 28.09 +/- 10.09
Iteration 104/500: rewards 33.66 +/- 20.14
Iteration 105/500: rewards 37.75 +/- 23.68
Iteration 106/500: rewards 31.47 +/- 17.36
Iteration 107/500: rewards 33.12 +/- 17.73
Iteration 108/500: rewards 33.28 +/- 20.83
Iteration 109/500: rewards 36.09 +/- 17.13
Iteration 110/500: rewards 40.06 +/- 21.72
Iteration 111/500: rewards 35.72 +/- 18.11
Iteration 112/500: rewards 34.0 +/- 25.06
Iteration 113/500: rewards 36.78 +/- 18.97
Iteration 114/500: rewards 42.25 +/- 24.46
Iteration 115/500: rewards 32.28 +/- 13.48
Iteration 116/500: rewards 40.19 +/- 20.48
Iteration 117/500: rewards 34.97 +/- 16.93
Iteration 118/500: rewards 39.06 +/- 22.99
Iteration 119/500: rewards 36.91 +/- 19.55
Iteration 120/500: rewards 35.69 +/- 19.11
Iteration 121/500: rewards 38.25 +/- 16.57
Iteration 122/500: rewards 31.59 +/- 12.6
Iteration 123/500: rewards 41.47 +/- 18.9
Iteration 124/500: rewards 43.72 +/- 23.42
Iteration 125/500: rewards 43.03 +/- 19.55
Iteration 126/500: rewards 42.0 +/- 19.01
Iteration 127/500: rewards 43.66 +/- 18.26
Iteration 128/500: rewards 48.72 +/- 26.73
Iteration 129/500: rewards 40.12 +/- 20.86
```

```
Iteration 130/500: rewards 47.06 +/- 23.14
Iteration 131/500: rewards 42.34 +/- 23.16
Iteration 132/500: rewards 40.0 +/- 15.99
Iteration 133/500: rewards 49.38 +/- 29.74
Iteration 134/500: rewards 51.88 +/- 31.54
Iteration 135/500: rewards 44.88 +/- 20.65
Iteration 136/500: rewards 46.62 +/- 18.09
Iteration 137/500: rewards 52.5 +/- 18.82
Iteration 138/500: rewards 51.31 +/- 23.59
Iteration 139/500: rewards 51.09 +/- 21.1
Iteration 140/500: rewards 46.84 +/- 15.58
Iteration 141/500: rewards 47.06 +/- 20.89
Iteration 142/500: rewards 45.41 +/- 22.63
Iteration 143/500: rewards 49.28 +/- 22.11
Iteration 144/500: rewards 57.78 +/- 20.81
Iteration 145/500: rewards 56.56 +/- 20.46
Iteration 146/500: rewards 50.97 +/- 23.36
Iteration 147/500: rewards 43.38 +/- 22.0
Iteration 148/500: rewards 48.94 +/- 19.98
Iteration 149/500: rewards 51.12 +/- 29.48
Iteration 150/500: rewards 52.78 +/- 20.43
Iteration 151/500: rewards 56.5 +/- 23.96
Iteration 152/500: rewards 45.44 +/- 14.36
Iteration 153/500: rewards 58.66 +/- 27.87
Iteration 154/500: rewards 58.41 +/- 29.32
Iteration 155/500: rewards 57.06 +/- 25.84
Iteration 156/500: rewards 63.09 +/- 22.32
Iteration 157/500: rewards 59.19 +/- 34.45
Iteration 158/500: rewards 58.97 +/- 33.7
Iteration 159/500: rewards 61.75 +/- 30.58
Iteration 160/500: rewards 55.09 +/- 22.78
Iteration 161/500: rewards 55.34 +/- 28.65
Iteration 162/500: rewards 59.09 +/- 28.62
Iteration 163/500: rewards 65.84 +/- 20.81
Iteration 164/500: rewards 59.38 +/- 28.56
Iteration 165/500: rewards 68.5 +/- 24.4
Iteration 166/500: rewards 67.0 +/- 28.38
Iteration 167/500: rewards 69.06 +/- 33.02
Iteration 168/500: rewards 56.34 +/- 22.34
Iteration 169/500: rewards 66.25 +/- 32.06
Iteration 170/500: rewards 67.5 +/- 23.44
Iteration 171/500: rewards 65.91 +/- 24.4
Iteration 172/500: rewards 62.97 +/- 18.78
Iteration 173/500: rewards 65.75 +/- 27.05
Iteration 174/500: rewards 69.44 +/- 18.14
Iteration 175/500: rewards 66.88 +/- 25.86
Iteration 176/500: rewards 75.78 +/- 41.72
Iteration 177/500: rewards 65.25 +/- 25.42
```

```
Iteration 178/500: rewards 70.53 +/- 21.74
Iteration 179/500: rewards 70.69 +/- 19.18
Iteration 180/500: rewards 60.62 +/- 19.06
Iteration 181/500: rewards 69.03 +/- 28.99
Iteration 182/500: rewards 69.94 +/- 32.5
Iteration 183/500: rewards 81.69 +/- 27.18
Iteration 184/500: rewards 68.41 +/- 28.08
Iteration 185/500: rewards 84.28 +/- 29.25
Iteration 186/500: rewards 67.69 +/- 28.27
Iteration 187/500: rewards 74.78 +/- 26.63
Iteration 188/500: rewards 70.91 +/- 26.63
Iteration 189/500: rewards 75.59 +/- 36.62
Iteration 190/500: rewards 75.28 +/- 26.73
Iteration 191/500: rewards 76.72 +/- 24.79
Iteration 192/500: rewards 72.53 +/- 18.81
Iteration 193/500: rewards 78.94 +/- 30.03
Iteration 194/500: rewards 81.41 +/- 32.96
Iteration 195/500: rewards 79.16 +/- 20.38
Iteration 196/500: rewards 66.19 +/- 25.36
Iteration 197/500: rewards 86.38 +/- 30.36
Iteration 198/500: rewards 74.31 +/- 28.22
Iteration 199/500: rewards 87.09 +/- 32.24
Iteration 200/500: rewards 82.12 +/- 36.95
Iteration 201/500: rewards 86.91 +/- 36.06
Iteration 202/500: rewards 87.03 +/- 31.61
Iteration 203/500: rewards 89.34 +/- 25.35
Iteration 204/500: rewards 91.78 +/- 39.63
Iteration 205/500: rewards 89.41 +/- 27.31
Iteration 206/500: rewards 88.62 +/- 27.96
Iteration 207/500: rewards 94.0 +/- 40.19
Iteration 208/500: rewards 91.59 +/- 34.06
Iteration 209/500: rewards 97.94 +/- 38.11
Iteration 210/500: rewards 88.62 +/- 31.22
Iteration 211/500: rewards 98.03 +/- 41.8
Iteration 212/500: rewards 99.88 +/- 29.49
Iteration 213/500: rewards 96.31 +/- 31.84
Iteration 214/500: rewards 99.44 +/- 37.86
Iteration 215/500: rewards 100.84 +/- 33.0
Iteration 216/500: rewards 109.94 +/- 42.75
Iteration 217/500: rewards 106.53 +/- 42.95
Iteration 218/500: rewards 100.44 +/- 40.27
Iteration 219/500: rewards 111.31 +/- 42.5
Iteration 220/500: rewards 118.0 +/- 28.23
Iteration 221/500: rewards 113.44 +/- 46.52
Iteration 222/500: rewards 113.31 +/- 29.04
Iteration 223/500: rewards 125.59 +/- 55.9
Iteration 224/500: rewards 102.47 +/- 27.18
Iteration 225/500: rewards 130.62 +/- 55.56
```

```
Iteration 226/500: rewards 125.81 +/- 55.26
Iteration 227/500: rewards 122.72 +/- 41.06
Iteration 228/500: rewards 144.72 +/- 60.08
Iteration 229/500: rewards 150.22 +/- 59.61
Iteration 230/500: rewards 154.66 +/- 65.31
Iteration 231/500: rewards 150.16 +/- 51.55
Iteration 232/500: rewards 166.97 +/- 59.14
Iteration 233/500: rewards 149.34 +/- 56.44
Iteration 234/500: rewards 154.56 +/- 58.25
Iteration 235/500: rewards 145.0 +/- 51.41
Iteration 236/500: rewards 146.59 +/- 46.4
Iteration 237/500: rewards 156.72 +/- 48.5
Iteration 238/500: rewards 169.47 +/- 62.06
Iteration 239/500: rewards 181.06 +/- 58.57
Iteration 240/500: rewards 179.53 +/- 74.14
Iteration 241/500: rewards 191.25 +/- 56.15
Iteration 242/500: rewards 193.84 +/- 63.17
Iteration 243/500: rewards 195.03 +/- 71.74
Iteration 244/500: rewards 216.69 +/- 69.51
Iteration 245/500: rewards 223.88 +/- 80.92
Iteration 246/500: rewards 201.5 +/- 66.94
Iteration 247/500: rewards 219.03 +/- 61.23
Iteration 248/500: rewards 224.97 +/- 70.38
Iteration 249/500: rewards 226.97 +/- 83.88
Iteration 250/500: rewards 214.44 +/- 57.1
Iteration 251/500: rewards 204.97 +/- 60.48
Iteration 252/500: rewards 232.75 +/- 79.25
Iteration 253/500: rewards 252.22 +/- 83.64
Iteration 254/500: rewards 250.75 +/- 86.56
Iteration 255/500: rewards 248.0 +/- 71.28
Iteration 256/500: rewards 260.38 +/- 80.59
Iteration 257/500: rewards 256.53 +/- 85.24
Iteration 258/500: rewards 250.72 +/- 91.65
Iteration 259/500: rewards 254.97 +/- 86.0
Iteration 260/500: rewards 286.0 +/- 94.87
Iteration 261/500: rewards 281.78 +/- 93.06
Iteration 262/500: rewards 296.31 +/- 75.67
Iteration 263/500: rewards 316.25 +/- 94.55
Iteration 264/500: rewards 332.38 +/- 119.66
Iteration 265/500: rewards 341.53 +/- 109.11
Iteration 266/500: rewards 337.5 +/- 119.96
Iteration 267/500: rewards 303.94 +/- 120.68
Iteration 268/500: rewards 272.38 +/- 122.36
Iteration 269/500: rewards 271.5 +/- 114.57
Iteration 270/500: rewards 268.12 +/- 109.87
Iteration 271/500: rewards 261.66 +/- 106.6
Iteration 272/500: rewards 249.5 +/- 81.12
Iteration 273/500: rewards 270.22 +/- 86.48
```

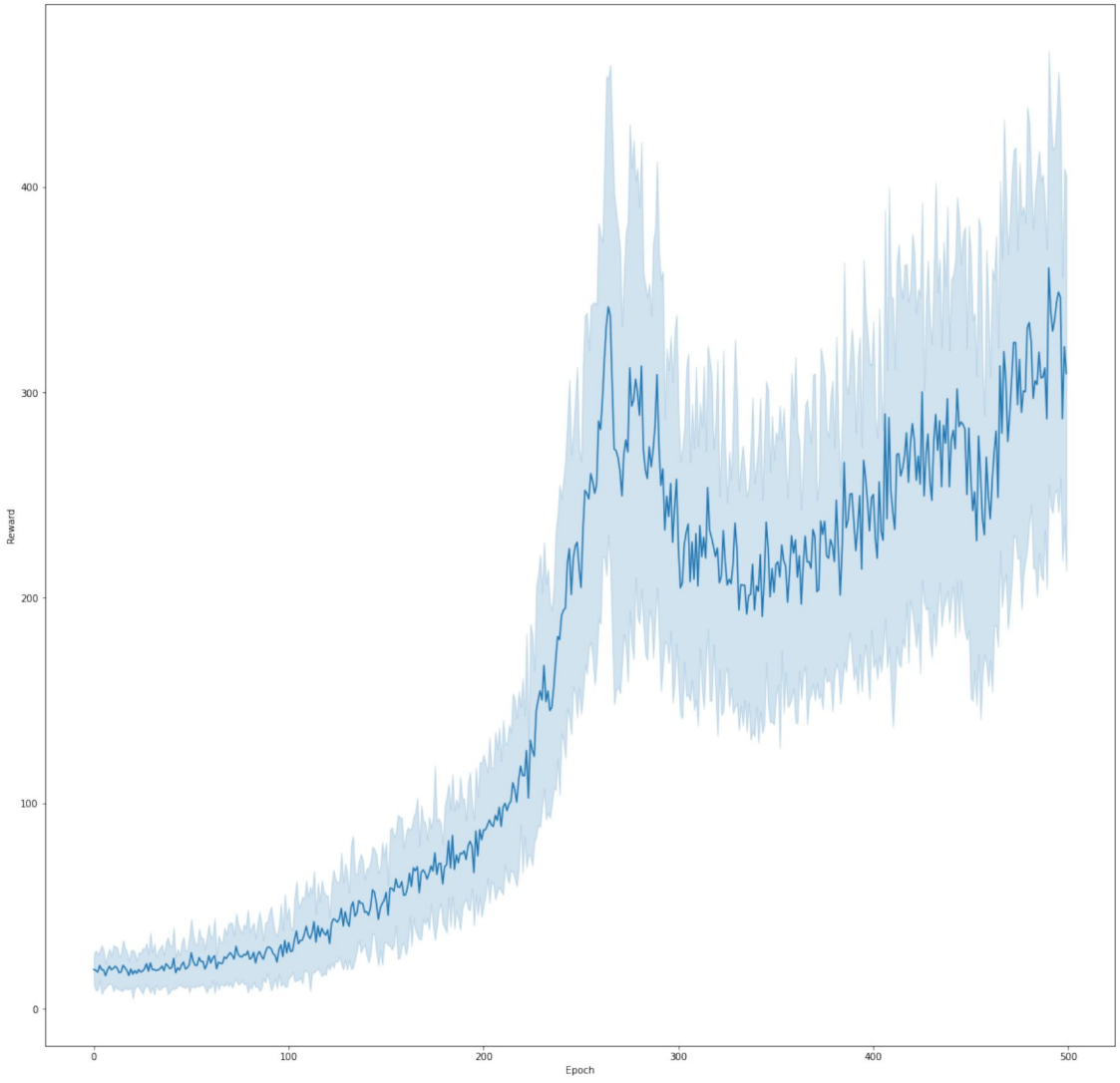
```
Iteration 274/500: rewards 276.78 +/- 99.39
Iteration 275/500: rewards 270.91 +/- 109.99
Iteration 276/500: rewards 311.84 +/- 116.61
Iteration 277/500: rewards 293.22 +/- 114.12
Iteration 278/500: rewards 296.22 +/- 124.25
Iteration 279/500: rewards 306.38 +/- 95.05
Iteration 280/500: rewards 299.81 +/- 107.45
Iteration 281/500: rewards 288.75 +/- 99.79
Iteration 282/500: rewards 312.72 +/- 107.19
Iteration 283/500: rewards 272.22 +/- 84.55
Iteration 284/500: rewards 262.34 +/- 88.33
Iteration 285/500: rewards 257.94 +/- 86.85
Iteration 286/500: rewards 273.44 +/- 78.16
Iteration 287/500: rewards 263.66 +/- 72.32
Iteration 288/500: rewards 271.84 +/- 99.63
Iteration 289/500: rewards 283.53 +/- 94.0
Iteration 290/500: rewards 308.56 +/- 102.17
Iteration 291/500: rewards 274.09 +/- 91.75
Iteration 292/500: rewards 254.62 +/- 98.5
Iteration 293/500: rewards 262.66 +/- 94.81
Iteration 294/500: rewards 232.94 +/- 53.29
Iteration 295/500: rewards 249.34 +/- 70.88
Iteration 296/500: rewards 239.44 +/- 70.25
Iteration 297/500: rewards 255.53 +/- 70.6
Iteration 298/500: rewards 226.97 +/- 76.81
Iteration 299/500: rewards 243.84 +/- 84.35
Iteration 300/500: rewards 257.59 +/- 78.75
Iteration 301/500: rewards 224.19 +/- 65.66
Iteration 302/500: rewards 204.78 +/- 60.68
Iteration 303/500: rewards 207.16 +/- 64.94
Iteration 304/500: rewards 225.91 +/- 56.42
Iteration 305/500: rewards 231.81 +/- 78.56
Iteration 306/500: rewards 235.81 +/- 81.6
Iteration 307/500: rewards 207.78 +/- 57.28
Iteration 308/500: rewards 227.09 +/- 66.18
Iteration 309/500: rewards 208.94 +/- 62.64
Iteration 310/500: rewards 231.19 +/- 79.72
Iteration 311/500: rewards 205.66 +/- 66.94
Iteration 312/500: rewards 235.06 +/- 58.84
Iteration 313/500: rewards 219.88 +/- 62.17
Iteration 314/500: rewards 229.44 +/- 82.08
Iteration 315/500: rewards 219.19 +/- 51.23
Iteration 316/500: rewards 253.59 +/- 67.82
Iteration 317/500: rewards 232.78 +/- 81.61
Iteration 318/500: rewards 229.19 +/- 77.98
Iteration 319/500: rewards 224.78 +/- 46.81
Iteration 320/500: rewards 219.91 +/- 64.17
Iteration 321/500: rewards 224.16 +/- 89.98
```

```
Iteration 322/500: rewards 207.19 +/- 41.41
Iteration 323/500: rewards 210.41 +/- 66.43
Iteration 324/500: rewards 232.66 +/- 86.42
Iteration 325/500: rewards 217.19 +/- 53.03
Iteration 326/500: rewards 206.09 +/- 39.78
Iteration 327/500: rewards 209.0 +/- 61.13
Iteration 328/500: rewards 206.72 +/- 49.73
Iteration 329/500: rewards 216.28 +/- 71.4
Iteration 330/500: rewards 236.28 +/- 87.94
Iteration 331/500: rewards 223.91 +/- 73.88
Iteration 332/500: rewards 193.94 +/- 57.26
Iteration 333/500: rewards 206.28 +/- 58.08
Iteration 334/500: rewards 206.03 +/- 67.01
Iteration 335/500: rewards 206.03 +/- 55.16
Iteration 336/500: rewards 192.0 +/- 56.37
Iteration 337/500: rewards 201.0 +/- 51.4
Iteration 338/500: rewards 201.56 +/- 69.63
Iteration 339/500: rewards 216.25 +/- 79.76
Iteration 340/500: rewards 193.97 +/- 60.45
Iteration 341/500: rewards 205.72 +/- 58.11
Iteration 342/500: rewards 203.12 +/- 72.45
Iteration 343/500: rewards 220.94 +/- 75.1
Iteration 344/500: rewards 190.81 +/- 55.77
Iteration 345/500: rewards 208.38 +/- 68.98
Iteration 346/500: rewards 236.75 +/- 67.48
Iteration 347/500: rewards 224.0 +/- 75.46
Iteration 348/500: rewards 200.38 +/- 60.02
Iteration 349/500: rewards 214.12 +/- 73.3
Iteration 350/500: rewards 202.53 +/- 63.35
Iteration 351/500: rewards 216.19 +/- 63.42
Iteration 352/500: rewards 217.44 +/- 58.79
Iteration 353/500: rewards 210.09 +/- 82.1
Iteration 354/500: rewards 225.59 +/- 50.59
Iteration 355/500: rewards 218.16 +/- 73.13
Iteration 356/500: rewards 215.22 +/- 50.96
Iteration 357/500: rewards 197.75 +/- 50.24
Iteration 358/500: rewards 212.44 +/- 63.22
Iteration 359/500: rewards 230.12 +/- 77.92
Iteration 360/500: rewards 221.72 +/- 64.78
Iteration 361/500: rewards 228.12 +/- 87.38
Iteration 362/500: rewards 209.91 +/- 69.9
Iteration 363/500: rewards 220.41 +/- 55.16
Iteration 364/500: rewards 196.81 +/- 45.64
Iteration 365/500: rewards 215.91 +/- 51.9
Iteration 366/500: rewards 229.94 +/- 62.51
Iteration 367/500: rewards 217.41 +/- 77.82
Iteration 368/500: rewards 217.5 +/- 65.77
Iteration 369/500: rewards 214.25 +/- 60.28
```

```
Iteration 370/500: rewards 233.19 +/- 73.97
Iteration 371/500: rewards 229.56 +/- 78.38
Iteration 372/500: rewards 202.88 +/- 43.66
Iteration 373/500: rewards 203.94 +/- 52.23
Iteration 374/500: rewards 237.28 +/- 82.84
Iteration 375/500: rewards 230.81 +/- 83.75
Iteration 376/500: rewards 237.22 +/- 70.04
Iteration 377/500: rewards 220.41 +/- 60.55
Iteration 378/500: rewards 219.16 +/- 58.18
Iteration 379/500: rewards 228.28 +/- 68.11
Iteration 380/500: rewards 225.16 +/- 79.2
Iteration 381/500: rewards 217.53 +/- 54.7
Iteration 382/500: rewards 247.41 +/- 78.45
Iteration 383/500: rewards 224.22 +/- 59.09
Iteration 384/500: rewards 201.25 +/- 52.03
Iteration 385/500: rewards 224.22 +/- 65.9
Iteration 386/500: rewards 265.81 +/- 95.75
Iteration 387/500: rewards 234.03 +/- 69.05
Iteration 388/500: rewards 237.59 +/- 60.67
Iteration 389/500: rewards 250.34 +/- 66.0
Iteration 390/500: rewards 250.66 +/- 78.81
Iteration 391/500: rewards 237.0 +/- 75.88
Iteration 392/500: rewards 222.88 +/- 56.32
Iteration 393/500: rewards 234.31 +/- 79.75
Iteration 394/500: rewards 249.62 +/- 75.95
Iteration 395/500: rewards 213.84 +/- 58.56
Iteration 396/500: rewards 266.88 +/- 96.01
Iteration 397/500: rewards 258.28 +/- 79.73
Iteration 398/500: rewards 246.31 +/- 81.11
Iteration 399/500: rewards 232.5 +/- 80.89
Iteration 400/500: rewards 248.59 +/- 63.35
Iteration 401/500: rewards 250.19 +/- 82.94
Iteration 402/500: rewards 228.91 +/- 58.02
Iteration 403/500: rewards 219.16 +/- 57.81
Iteration 404/500: rewards 256.38 +/- 83.13
Iteration 405/500: rewards 232.34 +/- 67.47
Iteration 406/500: rewards 227.94 +/- 52.96
Iteration 407/500: rewards 289.38 +/- 97.98
Iteration 408/500: rewards 238.34 +/- 71.18
Iteration 409/500: rewards 287.66 +/- 110.01
Iteration 410/500: rewards 251.69 +/- 93.61
Iteration 411/500: rewards 241.38 +/- 102.67
Iteration 412/500: rewards 233.19 +/- 76.83
Iteration 413/500: rewards 269.62 +/- 91.87
Iteration 414/500: rewards 270.09 +/- 100.39
Iteration 415/500: rewards 259.19 +/- 91.5
Iteration 416/500: rewards 262.75 +/- 81.57
Iteration 417/500: rewards 268.78 +/- 91.69
```

```
Iteration 418/500: rewards 280.22 +/- 80.91
Iteration 419/500: rewards 256.12 +/- 86.52
Iteration 420/500: rewards 274.34 +/- 74.01
Iteration 421/500: rewards 284.72 +/- 90.78
Iteration 422/500: rewards 276.84 +/- 90.18
Iteration 423/500: rewards 257.09 +/- 80.11
Iteration 424/500: rewards 268.84 +/- 80.13
Iteration 425/500: rewards 255.25 +/- 90.43
Iteration 426/500: rewards 300.19 +/- 90.59
Iteration 427/500: rewards 249.47 +/- 50.9
Iteration 428/500: rewards 269.12 +/- 73.9
Iteration 429/500: rewards 279.66 +/- 83.05
Iteration 430/500: rewards 257.03 +/- 76.88
Iteration 431/500: rewards 247.31 +/- 75.0
Iteration 432/500: rewards 275.78 +/- 82.27
Iteration 433/500: rewards 289.19 +/- 110.99
Iteration 434/500: rewards 271.69 +/- 75.65
Iteration 435/500: rewards 286.09 +/- 77.47
Iteration 436/500: rewards 254.0 +/- 66.88
Iteration 437/500: rewards 283.84 +/- 87.81
Iteration 438/500: rewards 275.16 +/- 75.59
Iteration 439/500: rewards 296.81 +/- 92.24
Iteration 440/500: rewards 253.97 +/- 65.11
Iteration 441/500: rewards 276.97 +/- 76.65
Iteration 442/500: rewards 281.44 +/- 74.46
Iteration 443/500: rewards 272.44 +/- 90.24
Iteration 444/500: rewards 301.53 +/- 91.88
Iteration 445/500: rewards 283.25 +/- 98.7
Iteration 446/500: rewards 285.53 +/- 75.86
Iteration 447/500: rewards 284.28 +/- 90.99
Iteration 448/500: rewards 282.03 +/- 96.89
Iteration 449/500: rewards 250.22 +/- 69.77
Iteration 450/500: rewards 282.56 +/- 96.78
Iteration 451/500: rewards 258.75 +/- 105.68
Iteration 452/500: rewards 242.41 +/- 91.13
Iteration 453/500: rewards 251.47 +/- 85.71
Iteration 454/500: rewards 227.69 +/- 78.96
Iteration 455/500: rewards 278.69 +/- 104.47
Iteration 456/500: rewards 260.09 +/- 117.36
Iteration 457/500: rewards 238.25 +/- 74.63
Iteration 458/500: rewards 230.56 +/- 57.3
Iteration 459/500: rewards 268.44 +/- 99.29
Iteration 460/500: rewards 249.38 +/- 90.95
Iteration 461/500: rewards 238.34 +/- 67.75
Iteration 462/500: rewards 257.19 +/- 101.43
Iteration 463/500: rewards 270.88 +/- 82.82
Iteration 464/500: rewards 281.09 +/- 93.52
Iteration 465/500: rewards 248.75 +/- 71.63
```

```
Iteration 466/500: rewards 312.81 +/- 88.59
Iteration 467/500: rewards 280.06 +/- 83.32
Iteration 468/500: rewards 319.78 +/- 111.28
Iteration 469/500: rewards 306.03 +/- 93.09
Iteration 470/500: rewards 275.97 +/- 89.58
Iteration 471/500: rewards 289.53 +/- 92.16
Iteration 472/500: rewards 306.53 +/- 96.43
Iteration 473/500: rewards 324.22 +/- 92.31
Iteration 474/500: rewards 324.12 +/- 93.47
Iteration 475/500: rewards 293.91 +/- 73.91
Iteration 476/500: rewards 315.94 +/- 94.46
Iteration 477/500: rewards 290.19 +/- 94.4
Iteration 478/500: rewards 300.59 +/- 88.49
Iteration 479/500: rewards 300.16 +/- 80.85
Iteration 480/500: rewards 331.38 +/- 105.54
Iteration 481/500: rewards 333.94 +/- 96.01
Iteration 482/500: rewards 324.53 +/- 78.0
Iteration 483/500: rewards 296.97 +/- 81.32
Iteration 484/500: rewards 305.59 +/- 89.83
Iteration 485/500: rewards 303.69 +/- 101.78
Iteration 486/500: rewards 319.5 +/- 96.46
Iteration 487/500: rewards 306.94 +/- 94.34
Iteration 488/500: rewards 307.25 +/- 97.2
Iteration 489/500: rewards 311.81 +/- 77.69
Iteration 490/500: rewards 287.03 +/- 81.56
Iteration 491/500: rewards 360.5 +/- 103.89
Iteration 492/500: rewards 339.91 +/- 94.08
Iteration 493/500: rewards 329.72 +/- 87.13
Iteration 494/500: rewards 334.94 +/- 82.92
Iteration 495/500: rewards 343.59 +/- 89.55
Iteration 496/500: rewards 348.72 +/- 105.48
Iteration 497/500: rewards 346.09 +/- 86.55
Iteration 498/500: rewards 287.09 +/- 67.93
Iteration 499/500: rewards 322.16 +/- 85.51
Iteration 500/500: rewards 309.03 +/- 94.53
```



[ ]:
```
# You will be graded on this output of this cell; so kindly run it
agent.evaluate()
```

```
<IPython.core.display.HTML object>
```

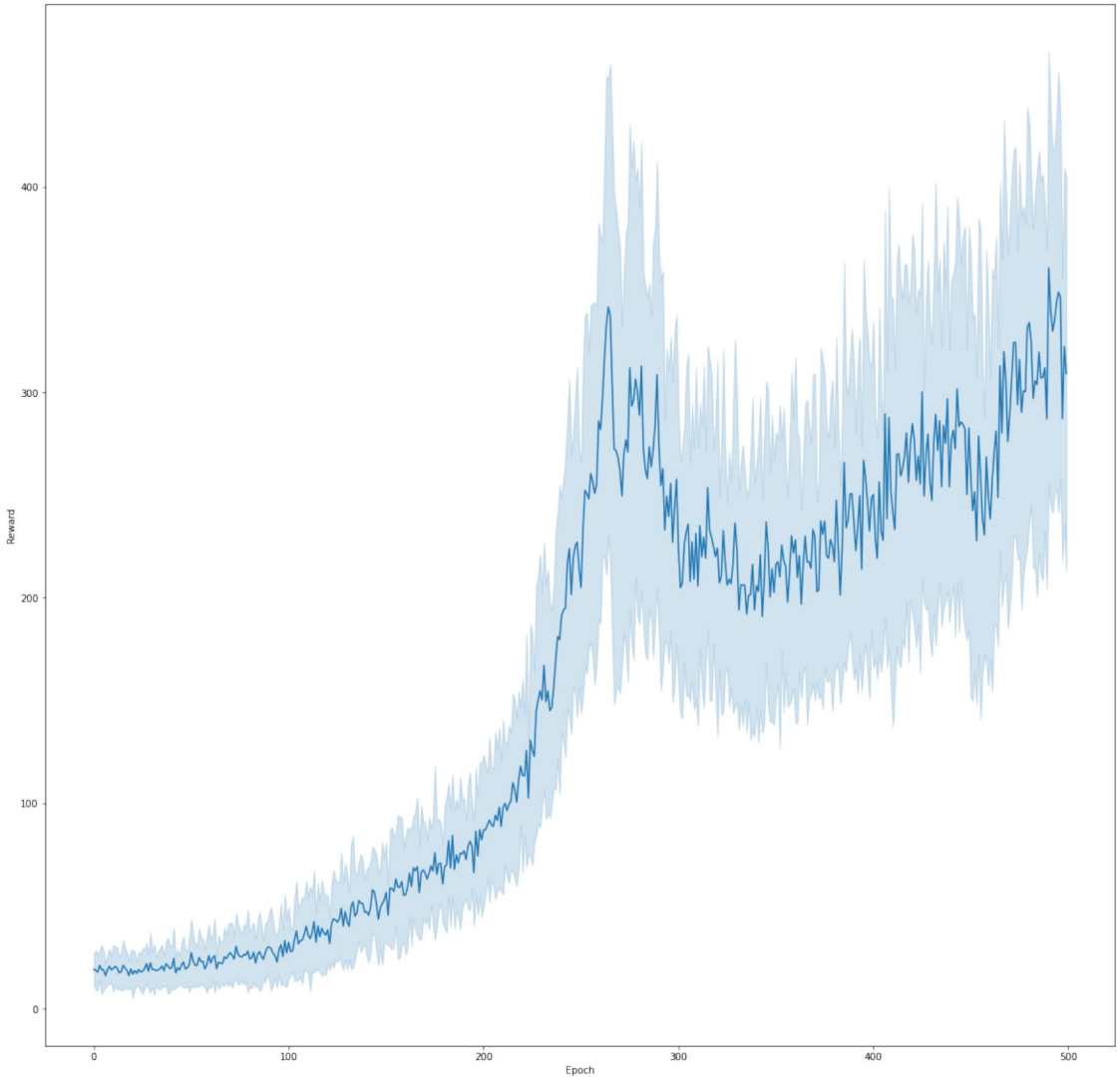
```
Reward: 500.0
```

5.0.2 Qn 2.2: Eventhough the previous `REINFORCEv2+B` agent used a value estimator network similar to that of the Actor-Critic agent why is not called an Actor-Critic method ? [3 Marks]

Reinforcev2+B is not called Actor-Critic because the output of the Critic is not used in bootstrapping and only used as a baseline.

4.1 Qn 2.1 Implement a one-step Actor-Critic agent **15 / 15**

✓ **- 0 pts** correct



[ ]:
```
# You will be graded on this output of this cell; so kindly run it
agent.evaluate()
```

```
<IPython.core.display.HTML object>
```

```
Reward: 500.0
```

5.0.2 Qn 2.2: Eventhough the previous `REINFORCEv2+B` agent used a value estimator network similar to that of the Actor-Critic agent why is not called an Actor-Critic method ? [3 Marks]

Reinforcev2+B is not called Actor-Critic because the output of the Critic is not used in bootstrapping and only used as a baseline.

4.2 Qn 2.2: Eventhough the previous REINFORCEv2+B agent used a value estimator network similar to that of the Actor-Critic agent why is not called an Actor-Critic method ?

2 / 3

✓ **- 1 pts** **insufficient explanation**

5.0.3 Qn 2.3: How does the Actor-Critic algorithm reduces variance? What about bias? We are using one-step rewards here, is there a way we can strike a balance between variance and bias? [5 Marks]

The AC algorithm reduces variance by approximating the expected return (in our case using a neural network) instead of of using a MC sample computation, this will let us use "the one-step return which is often superior to the actual return in terms of its variance and computational congeniality, even though it introduces bias." (source: the book) the balance between bias and variance is a trade-off that we always see in Machine Learning. in Reinforcment Learning, we can strike a balance by using n-step returns and eligibility traces. (where n depends on the problem)

5.0.4 Qn 2.4: Challenge! Can you tweak the hyperparameters of Actor-Critic to achieve better performance? Compare your results againts what you already have in section 3.1, in a single plot. [5 Marks]

Tune γ within the same range as in Qn. 1.3.c and tune the hypereparameters of the value networks.

[ ]:
```
# Insert your code here to search for best hyper-parameters
gammas = [1, 0.99, 0.95]
values_layers = [16, 8]
policy_learning_rates = [1e-2, 1e-3, 1e-4]
value_learning_rates = [1e-2, 1e-3, 1e-4]




def plot_everything(dic):
  fig, ax = plt.subplots()
  for k in list(dic.keys()):

    BaseAgent.plot_rewards(dic[k], ax)

  plt.rcParams['figure.figsize'] = [12, 12]
  plt.legend(labels=list(dic.keys()))



def test_AC(gamma, v_lr, p_lr):
  config = {
    'env_id': 'CartPole-v1',
    'seed': 8953,
    'gamma': gamma,
    'policy_layers': [16, 8],
    'policy_learning_rate': p_lr,
    'use_baseline': True,
    'value_layers': [16, 8],
    'value_learning_rate': v_lr,
  }
```

4.3 Qn 2.3: How does the Actor-Critic algorithm reduces variance? What about bias? We are using one-step rewards here, is there a way we can strike a balance between variance and bias? **5 / 5**

✓ **- 0 pts** Correct

5.0.3 Qn 2.3: How does the Actor-Critic algorithm reduces variance? What about bias? We are using one-step rewards here, is there a way we can strike a balance between variance and bias? [5 Marks]

The AC algorithm reduces variance by approximating the expected return (in our case using a neural network) instead of of using a MC sample computation, this will let us use "the one-step return which is often superior to the actual return in terms of its variance and computational congeniality, even though it introduces bias." (source: the book) the balance between bias and variance is a trade-off that we always see in Machine Learning. in Reinforcment Learning, we can strike a balance by using n-step returns and eligibility traces. (where n depends on the problem)

5.0.4 Qn 2.4: Challenge! Can you tweak the hyperparameters of Actor-Critic to achieve better performance? Compare your results againts what you already have in section 3.1, in a single plot. [5 Marks]

Tune γ within the same range as in Qn. 1.3.c and tune the hypereparameters of the value networks.

[ ]:
```
# Insert your code here to search for best hyper-parameters
gammas = [1, 0.99, 0.95]
values_layers = [16, 8]
policy_learning_rates = [1e-2, 1e-3, 1e-4]
value_learning_rates = [1e-2, 1e-3, 1e-4]




def plot_everything(dic):
  fig, ax = plt.subplots()
  for k in list(dic.keys()):

    BaseAgent.plot_rewards(dic[k], ax)

  plt.rcParams['figure.figsize'] = [12, 12]
  plt.legend(labels=list(dic.keys()))



def test_AC(gamma, v_lr, p_lr):
  config = {
    'env_id': 'CartPole-v1',
    'seed': 8953,
    'gamma': gamma,
    'policy_layers': [16, 8],
    'policy_learning_rate': p_lr,
    'use_baseline': True,
    'value_layers': [16, 8],
    'value_learning_rate': v_lr,
  }
```

```
  agent = ActorCriticAgent(config)
  print('_'*50)
  print('The gamma chosen is: {}'.format(gamma))
  print('The value lr chosen is {}'.format(v_lr))
  print('The policy lr chosen is {}'.format(p_lr))
  ActorCritic_rewards2 = agent.train(n_episodes=32, n_iterations=500)
  avg = np.mean(ActorCritic_rewards2[-50:] )
  print('The average reward is {}'.format(avg))
  return avg, ActorCritic_rewards2


avg_rewards = {}
rewards = {}

for g in gammas:
  for p in policy_learning_rates:
    for v in value_learning_rates:

      avg, full_rewards = test_AC(g, v,p)
      avg_rewards['{}_{}_{}'.format(g,p,v)] = avg
      rewards['{}_{}_{}'.format(g,p,v)] = full_rewards
```

```
Streaming output truncated to the last 5000 lines.
Iteration 56/500: rewards 20.69 +/- 10.18
Iteration 57/500: rewards 21.5 +/- 11.61
Iteration 58/500: rewards 19.84 +/- 8.64
Iteration 59/500: rewards 17.59 +/- 7.7
Iteration 60/500: rewards 17.22 +/- 8.44
Iteration 61/500: rewards 22.03 +/- 12.13
Iteration 62/500: rewards 20.84 +/- 8.16
Iteration 63/500: rewards 20.28 +/- 13.25
Iteration 64/500: rewards 21.0 +/- 10.61
Iteration 65/500: rewards 17.78 +/- 6.45
Iteration 66/500: rewards 17.12 +/- 7.02
Iteration 67/500: rewards 17.16 +/- 6.36
Iteration 68/500: rewards 16.22 +/- 6.22
Iteration 69/500: rewards 18.0 +/- 10.11
Iteration 70/500: rewards 19.41 +/- 10.13
```

```
Iteration 71/500: rewards 16.72 +/- 6.3
Iteration 72/500: rewards 22.97 +/- 15.34
Iteration 73/500: rewards 19.34 +/- 10.74
Iteration 74/500: rewards 19.81 +/- 9.58
Iteration 75/500: rewards 17.09 +/- 6.34
Iteration 76/500: rewards 18.94 +/- 7.44
Iteration 77/500: rewards 20.69 +/- 10.11
Iteration 78/500: rewards 20.19 +/- 8.57
Iteration 79/500: rewards 20.47 +/- 9.76
Iteration 80/500: rewards 22.75 +/- 12.31
Iteration 81/500: rewards 17.03 +/- 6.47
Iteration 82/500: rewards 18.59 +/- 7.31
Iteration 83/500: rewards 17.66 +/- 8.48
Iteration 84/500: rewards 21.06 +/- 14.14
Iteration 85/500: rewards 15.34 +/- 6.04
Iteration 86/500: rewards 19.44 +/- 8.34
Iteration 87/500: rewards 21.34 +/- 10.82
Iteration 88/500: rewards 17.94 +/- 7.67
Iteration 89/500: rewards 20.22 +/- 12.11
Iteration 90/500: rewards 20.41 +/- 9.84
Iteration 91/500: rewards 18.94 +/- 7.79
Iteration 92/500: rewards 21.53 +/- 8.79
Iteration 93/500: rewards 19.0 +/- 9.12
Iteration 94/500: rewards 20.22 +/- 10.88
Iteration 95/500: rewards 19.41 +/- 12.41
Iteration 96/500: rewards 19.66 +/- 9.6
Iteration 97/500: rewards 18.69 +/- 7.09
Iteration 98/500: rewards 18.88 +/- 9.27
Iteration 99/500: rewards 21.34 +/- 10.16
Iteration 100/500: rewards 19.31 +/- 9.82
Iteration 101/500: rewards 21.53 +/- 11.33
Iteration 102/500: rewards 19.66 +/- 9.94
Iteration 103/500: rewards 17.94 +/- 7.93
Iteration 104/500: rewards 22.0 +/- 12.8
Iteration 105/500: rewards 22.38 +/- 13.67
Iteration 106/500: rewards 18.5 +/- 8.81
Iteration 107/500: rewards 23.28 +/- 14.29
Iteration 108/500: rewards 18.19 +/- 7.8
Iteration 109/500: rewards 17.16 +/- 7.56
Iteration 110/500: rewards 17.25 +/- 6.58
Iteration 111/500: rewards 21.84 +/- 11.58
Iteration 112/500: rewards 17.94 +/- 8.02
Iteration 113/500: rewards 19.22 +/- 9.32
Iteration 114/500: rewards 20.31 +/- 11.16
Iteration 115/500: rewards 18.72 +/- 8.46
Iteration 116/500: rewards 19.0 +/- 8.56
Iteration 117/500: rewards 22.59 +/- 12.55
Iteration 118/500: rewards 18.28 +/- 6.91
```

```
Iteration 119/500: rewards 18.78 +/- 8.93
Iteration 120/500: rewards 19.56 +/- 9.55
Iteration 121/500: rewards 15.97 +/- 6.04
Iteration 122/500: rewards 21.03 +/- 11.35
Iteration 123/500: rewards 20.5 +/- 10.66
Iteration 124/500: rewards 20.84 +/- 8.3
Iteration 125/500: rewards 22.0 +/- 12.15
Iteration 126/500: rewards 19.91 +/- 10.37
Iteration 127/500: rewards 20.22 +/- 12.22
Iteration 128/500: rewards 20.31 +/- 9.93
Iteration 129/500: rewards 17.12 +/- 7.41
Iteration 130/500: rewards 20.0 +/- 10.6
Iteration 131/500: rewards 21.12 +/- 9.18
Iteration 132/500: rewards 23.03 +/- 13.03
Iteration 133/500: rewards 17.25 +/- 6.34
Iteration 134/500: rewards 17.38 +/- 7.36
Iteration 135/500: rewards 21.19 +/- 9.91
Iteration 136/500: rewards 18.72 +/- 6.02
Iteration 137/500: rewards 21.75 +/- 12.93
Iteration 138/500: rewards 21.34 +/- 8.74
Iteration 139/500: rewards 18.44 +/- 9.93
Iteration 140/500: rewards 19.84 +/- 11.07
Iteration 141/500: rewards 20.06 +/- 9.11
Iteration 142/500: rewards 18.97 +/- 9.04
Iteration 143/500: rewards 18.25 +/- 6.95
Iteration 144/500: rewards 20.03 +/- 10.54
Iteration 145/500: rewards 18.72 +/- 9.3
Iteration 146/500: rewards 23.03 +/- 11.29
Iteration 147/500: rewards 16.75 +/- 6.22
Iteration 148/500: rewards 20.72 +/- 12.46
Iteration 149/500: rewards 18.06 +/- 7.71
Iteration 150/500: rewards 20.97 +/- 10.92
Iteration 151/500: rewards 16.47 +/- 6.55
Iteration 152/500: rewards 18.44 +/- 7.6
Iteration 153/500: rewards 18.62 +/- 7.36
Iteration 154/500: rewards 21.28 +/- 13.21
Iteration 155/500: rewards 19.53 +/- 10.58
Iteration 156/500: rewards 17.28 +/- 5.13
Iteration 157/500: rewards 20.03 +/- 7.61
Iteration 158/500: rewards 19.59 +/- 11.67
Iteration 159/500: rewards 21.28 +/- 10.23
Iteration 160/500: rewards 21.5 +/- 13.63
Iteration 161/500: rewards 17.94 +/- 7.71
Iteration 162/500: rewards 24.44 +/- 14.53
Iteration 163/500: rewards 18.69 +/- 5.72
Iteration 164/500: rewards 20.0 +/- 12.91
Iteration 165/500: rewards 20.75 +/- 7.62
Iteration 166/500: rewards 19.06 +/- 8.92
```

```
Iteration 167/500: rewards 20.16 +/- 9.77
Iteration 168/500: rewards 20.5 +/- 10.15
Iteration 169/500: rewards 19.56 +/- 9.11
Iteration 170/500: rewards 17.91 +/- 4.98
Iteration 171/500: rewards 18.72 +/- 7.69
Iteration 172/500: rewards 19.69 +/- 11.14
Iteration 173/500: rewards 20.44 +/- 6.8
Iteration 174/500: rewards 18.97 +/- 9.36
Iteration 175/500: rewards 19.69 +/- 9.07
Iteration 176/500: rewards 20.59 +/- 10.76
Iteration 177/500: rewards 22.19 +/- 12.53
Iteration 178/500: rewards 22.62 +/- 12.1
Iteration 179/500: rewards 20.0 +/- 10.05
Iteration 180/500: rewards 20.94 +/- 11.43
Iteration 181/500: rewards 20.22 +/- 12.82
Iteration 182/500: rewards 20.59 +/- 13.16
Iteration 183/500: rewards 18.41 +/- 6.5
Iteration 184/500: rewards 19.69 +/- 9.0
Iteration 185/500: rewards 19.12 +/- 8.03
Iteration 186/500: rewards 16.78 +/- 5.88
Iteration 187/500: rewards 18.03 +/- 7.58
Iteration 188/500: rewards 23.75 +/- 11.83
Iteration 189/500: rewards 20.38 +/- 11.39
Iteration 190/500: rewards 18.34 +/- 7.88
Iteration 191/500: rewards 21.69 +/- 11.33
Iteration 192/500: rewards 20.25 +/- 8.01
Iteration 193/500: rewards 20.47 +/- 10.09
Iteration 194/500: rewards 19.22 +/- 6.09
Iteration 195/500: rewards 20.41 +/- 10.64
Iteration 196/500: rewards 18.94 +/- 9.41
Iteration 197/500: rewards 19.09 +/- 8.57
Iteration 198/500: rewards 20.31 +/- 7.21
Iteration 199/500: rewards 22.56 +/- 12.17
Iteration 200/500: rewards 18.81 +/- 10.1
Iteration 201/500: rewards 18.94 +/- 7.41
Iteration 202/500: rewards 18.0 +/- 9.45
Iteration 203/500: rewards 21.0 +/- 9.16
Iteration 204/500: rewards 21.0 +/- 11.98
Iteration 205/500: rewards 22.38 +/- 10.6
Iteration 206/500: rewards 19.5 +/- 8.58
Iteration 207/500: rewards 20.53 +/- 13.21
Iteration 208/500: rewards 24.53 +/- 13.83
Iteration 209/500: rewards 19.06 +/- 9.89
Iteration 210/500: rewards 20.47 +/- 9.61
Iteration 211/500: rewards 20.12 +/- 7.51
Iteration 212/500: rewards 19.75 +/- 11.72
Iteration 213/500: rewards 20.44 +/- 9.58
Iteration 214/500: rewards 17.62 +/- 8.55
```

```
Iteration 215/500: rewards 25.16 +/- 16.34
Iteration 216/500: rewards 19.19 +/- 12.59
Iteration 217/500: rewards 15.53 +/- 6.44
Iteration 218/500: rewards 19.41 +/- 10.11
Iteration 219/500: rewards 20.69 +/- 7.82
Iteration 220/500: rewards 20.81 +/- 9.4
Iteration 221/500: rewards 19.38 +/- 8.63
Iteration 222/500: rewards 20.88 +/- 10.65
Iteration 223/500: rewards 19.41 +/- 8.91
Iteration 224/500: rewards 20.03 +/- 10.24
Iteration 225/500: rewards 18.0 +/- 8.26
Iteration 226/500: rewards 19.56 +/- 7.71
Iteration 227/500: rewards 17.28 +/- 7.2
Iteration 228/500: rewards 17.94 +/- 6.42
Iteration 229/500: rewards 21.44 +/- 11.37
Iteration 230/500: rewards 18.53 +/- 8.58
Iteration 231/500: rewards 20.88 +/- 9.88
Iteration 232/500: rewards 21.97 +/- 10.53
Iteration 233/500: rewards 18.81 +/- 9.27
Iteration 234/500: rewards 17.09 +/- 6.76
Iteration 235/500: rewards 18.38 +/- 11.03
Iteration 236/500: rewards 19.25 +/- 6.81
Iteration 237/500: rewards 20.0 +/- 8.67
Iteration 238/500: rewards 18.38 +/- 8.26
Iteration 239/500: rewards 21.5 +/- 9.95
Iteration 240/500: rewards 18.81 +/- 9.28
Iteration 241/500: rewards 17.69 +/- 8.16
Iteration 242/500: rewards 21.56 +/- 15.34
Iteration 243/500: rewards 19.72 +/- 11.87
Iteration 244/500: rewards 18.66 +/- 8.47
Iteration 245/500: rewards 18.16 +/- 6.65
Iteration 246/500: rewards 19.22 +/- 9.35
Iteration 247/500: rewards 21.22 +/- 8.23
Iteration 248/500: rewards 21.12 +/- 11.88
Iteration 249/500: rewards 20.88 +/- 9.17
Iteration 250/500: rewards 20.25 +/- 9.29
Iteration 251/500: rewards 20.47 +/- 9.71
Iteration 252/500: rewards 17.47 +/- 7.65
Iteration 253/500: rewards 19.31 +/- 9.11
Iteration 254/500: rewards 21.84 +/- 8.57
Iteration 255/500: rewards 20.34 +/- 8.83
Iteration 256/500: rewards 17.72 +/- 5.73
Iteration 257/500: rewards 16.03 +/- 4.97
Iteration 258/500: rewards 22.75 +/- 11.23
Iteration 259/500: rewards 22.69 +/- 13.44
Iteration 260/500: rewards 19.78 +/- 11.17
Iteration 261/500: rewards 20.28 +/- 10.33
Iteration 262/500: rewards 18.94 +/- 8.89
```

```
Iteration 263/500: rewards 18.38 +/- 8.16
Iteration 264/500: rewards 16.94 +/- 6.02
Iteration 265/500: rewards 16.56 +/- 8.54
Iteration 266/500: rewards 19.88 +/- 9.98
Iteration 267/500: rewards 17.22 +/- 6.79
Iteration 268/500: rewards 22.03 +/- 10.94
Iteration 269/500: rewards 21.44 +/- 12.03
Iteration 270/500: rewards 22.19 +/- 13.19
Iteration 271/500: rewards 18.59 +/- 7.99
Iteration 272/500: rewards 20.06 +/- 11.32
Iteration 273/500: rewards 17.47 +/- 5.74
Iteration 274/500: rewards 18.5 +/- 7.88
Iteration 275/500: rewards 22.41 +/- 17.23
Iteration 276/500: rewards 20.5 +/- 9.76
Iteration 277/500: rewards 20.0 +/- 10.47
Iteration 278/500: rewards 21.69 +/- 9.72
Iteration 279/500: rewards 16.81 +/- 5.7
Iteration 280/500: rewards 18.06 +/- 6.74
Iteration 281/500: rewards 21.59 +/- 12.1
Iteration 282/500: rewards 18.94 +/- 7.53
Iteration 283/500: rewards 18.5 +/- 6.93
Iteration 284/500: rewards 19.69 +/- 7.11
Iteration 285/500: rewards 16.09 +/- 4.47
Iteration 286/500: rewards 21.06 +/- 11.5
Iteration 287/500: rewards 20.69 +/- 10.3
Iteration 288/500: rewards 19.41 +/- 11.53
Iteration 289/500: rewards 18.53 +/- 7.24
Iteration 290/500: rewards 19.72 +/- 9.6
Iteration 291/500: rewards 22.06 +/- 8.79
Iteration 292/500: rewards 19.81 +/- 10.47
Iteration 293/500: rewards 22.12 +/- 16.17
Iteration 294/500: rewards 20.97 +/- 9.3
Iteration 295/500: rewards 20.31 +/- 6.85
Iteration 296/500: rewards 19.44 +/- 8.27
Iteration 297/500: rewards 19.28 +/- 10.89
Iteration 298/500: rewards 20.5 +/- 10.33
Iteration 299/500: rewards 20.31 +/- 11.25
Iteration 300/500: rewards 16.0 +/- 5.17
Iteration 301/500: rewards 20.72 +/- 11.81
Iteration 302/500: rewards 19.09 +/- 7.64
Iteration 303/500: rewards 18.44 +/- 7.76
Iteration 304/500: rewards 21.44 +/- 10.7
Iteration 305/500: rewards 18.0 +/- 9.79
Iteration 306/500: rewards 19.78 +/- 7.75
Iteration 307/500: rewards 18.38 +/- 8.93
Iteration 308/500: rewards 22.03 +/- 12.26
Iteration 309/500: rewards 18.03 +/- 5.43
Iteration 310/500: rewards 16.97 +/- 6.0
```

```
Iteration 311/500: rewards 19.91 +/- 11.44
Iteration 312/500: rewards 18.75 +/- 7.08
Iteration 313/500: rewards 21.44 +/- 13.8
Iteration 314/500: rewards 18.53 +/- 10.32
Iteration 315/500: rewards 18.41 +/- 7.66
Iteration 316/500: rewards 21.53 +/- 12.26
Iteration 317/500: rewards 17.84 +/- 6.03
Iteration 318/500: rewards 18.06 +/- 6.03
Iteration 319/500: rewards 20.25 +/- 11.46
Iteration 320/500: rewards 20.62 +/- 11.31
Iteration 321/500: rewards 21.44 +/- 8.64
Iteration 322/500: rewards 17.53 +/- 10.71
Iteration 323/500: rewards 18.41 +/- 8.07
Iteration 324/500: rewards 19.66 +/- 8.07
Iteration 325/500: rewards 19.59 +/- 8.1
Iteration 326/500: rewards 18.75 +/- 6.73
Iteration 327/500: rewards 20.16 +/- 11.09
Iteration 328/500: rewards 18.72 +/- 11.69
Iteration 329/500: rewards 19.94 +/- 9.33
Iteration 330/500: rewards 22.75 +/- 9.9
Iteration 331/500: rewards 21.31 +/- 11.41
Iteration 332/500: rewards 21.03 +/- 16.11
Iteration 333/500: rewards 17.69 +/- 6.17
Iteration 334/500: rewards 20.0 +/- 7.43
Iteration 335/500: rewards 20.03 +/- 8.01
Iteration 336/500: rewards 17.62 +/- 6.85
Iteration 337/500: rewards 19.62 +/- 10.47
Iteration 338/500: rewards 18.06 +/- 8.05
Iteration 339/500: rewards 18.91 +/- 7.58
Iteration 340/500: rewards 17.91 +/- 8.8
Iteration 341/500: rewards 20.69 +/- 10.33
Iteration 342/500: rewards 18.88 +/- 8.08
Iteration 343/500: rewards 18.34 +/- 9.45
Iteration 344/500: rewards 18.97 +/- 10.27
Iteration 345/500: rewards 20.25 +/- 8.44
Iteration 346/500: rewards 19.16 +/- 9.1
Iteration 347/500: rewards 18.75 +/- 10.93
Iteration 348/500: rewards 21.0 +/- 12.29
Iteration 349/500: rewards 24.25 +/- 15.24
Iteration 350/500: rewards 18.28 +/- 6.99
Iteration 351/500: rewards 18.97 +/- 8.71
Iteration 352/500: rewards 20.31 +/- 10.78
Iteration 353/500: rewards 21.06 +/- 10.37
Iteration 354/500: rewards 20.16 +/- 10.05
Iteration 355/500: rewards 17.94 +/- 7.35
Iteration 356/500: rewards 18.91 +/- 10.71
Iteration 357/500: rewards 18.97 +/- 8.18
Iteration 358/500: rewards 19.06 +/- 9.72
```

```
Iteration 359/500: rewards 20.19 +/- 9.06
Iteration 360/500: rewards 20.12 +/- 13.29
Iteration 361/500: rewards 17.91 +/- 8.12
Iteration 362/500: rewards 21.53 +/- 8.87
Iteration 363/500: rewards 16.53 +/- 5.22
Iteration 364/500: rewards 19.84 +/- 11.04
Iteration 365/500: rewards 16.78 +/- 6.62
Iteration 366/500: rewards 21.25 +/- 13.36
Iteration 367/500: rewards 18.69 +/- 8.67
Iteration 368/500: rewards 19.41 +/- 8.38
Iteration 369/500: rewards 20.28 +/- 13.22
Iteration 370/500: rewards 17.56 +/- 9.43
Iteration 371/500: rewards 21.66 +/- 7.95
Iteration 372/500: rewards 22.19 +/- 9.86
Iteration 373/500: rewards 21.94 +/- 10.92
Iteration 374/500: rewards 19.03 +/- 7.2
Iteration 375/500: rewards 22.62 +/- 12.78
Iteration 376/500: rewards 16.31 +/- 5.5
Iteration 377/500: rewards 18.91 +/- 10.86
Iteration 378/500: rewards 19.34 +/- 9.31
Iteration 379/500: rewards 17.25 +/- 8.89
Iteration 380/500: rewards 19.66 +/- 8.55
Iteration 381/500: rewards 19.12 +/- 9.69
Iteration 382/500: rewards 18.84 +/- 8.39
Iteration 383/500: rewards 19.66 +/- 8.67
Iteration 384/500: rewards 22.53 +/- 13.92
Iteration 385/500: rewards 17.41 +/- 7.25
Iteration 386/500: rewards 20.19 +/- 8.01
Iteration 387/500: rewards 18.62 +/- 8.22
Iteration 388/500: rewards 17.22 +/- 7.26
Iteration 389/500: rewards 19.69 +/- 8.48
Iteration 390/500: rewards 16.81 +/- 5.9
Iteration 391/500: rewards 20.69 +/- 9.84
Iteration 392/500: rewards 18.5 +/- 9.66
Iteration 393/500: rewards 17.5 +/- 6.52
Iteration 394/500: rewards 17.97 +/- 6.87
Iteration 395/500: rewards 20.19 +/- 7.54
Iteration 396/500: rewards 20.16 +/- 11.11
Iteration 397/500: rewards 19.81 +/- 10.93
Iteration 398/500: rewards 22.47 +/- 10.44
Iteration 399/500: rewards 21.25 +/- 10.72
Iteration 400/500: rewards 22.81 +/- 11.57
Iteration 401/500: rewards 19.38 +/- 9.84
Iteration 402/500: rewards 16.16 +/- 5.44
Iteration 403/500: rewards 20.75 +/- 11.98
Iteration 404/500: rewards 18.72 +/- 10.53
Iteration 405/500: rewards 19.53 +/- 8.05
Iteration 406/500: rewards 20.25 +/- 9.81
```

```
Iteration 407/500: rewards 20.53 +/- 11.04
Iteration 408/500: rewards 19.19 +/- 9.12
Iteration 409/500: rewards 17.66 +/- 8.31
Iteration 410/500: rewards 18.69 +/- 9.5
Iteration 411/500: rewards 20.34 +/- 9.7
Iteration 412/500: rewards 20.56 +/- 11.45
Iteration 413/500: rewards 19.59 +/- 11.37
Iteration 414/500: rewards 20.12 +/- 9.35
Iteration 415/500: rewards 17.97 +/- 8.19
Iteration 416/500: rewards 19.53 +/- 9.37
Iteration 417/500: rewards 19.97 +/- 10.55
Iteration 418/500: rewards 16.78 +/- 7.25
Iteration 419/500: rewards 20.81 +/- 8.48
Iteration 420/500: rewards 22.38 +/- 10.95
Iteration 421/500: rewards 18.06 +/- 7.39
Iteration 422/500: rewards 19.34 +/- 10.85
Iteration 423/500: rewards 21.59 +/- 9.63
Iteration 424/500: rewards 22.0 +/- 11.75
Iteration 425/500: rewards 19.28 +/- 8.94
Iteration 426/500: rewards 21.16 +/- 8.73
Iteration 427/500: rewards 20.78 +/- 9.54
Iteration 428/500: rewards 20.84 +/- 10.61
Iteration 429/500: rewards 20.09 +/- 10.6
Iteration 430/500: rewards 19.09 +/- 8.24
Iteration 431/500: rewards 15.81 +/- 5.53
Iteration 432/500: rewards 16.62 +/- 5.38
Iteration 433/500: rewards 21.56 +/- 14.24
Iteration 434/500: rewards 19.19 +/- 9.91
Iteration 435/500: rewards 18.69 +/- 9.57
Iteration 436/500: rewards 19.72 +/- 8.09
Iteration 437/500: rewards 20.06 +/- 11.63
Iteration 438/500: rewards 21.91 +/- 16.52
Iteration 439/500: rewards 20.03 +/- 9.35
Iteration 440/500: rewards 20.56 +/- 9.59
Iteration 441/500: rewards 20.88 +/- 10.91
Iteration 442/500: rewards 17.31 +/- 7.81
Iteration 443/500: rewards 18.81 +/- 9.19
Iteration 444/500: rewards 20.5 +/- 10.4
Iteration 445/500: rewards 20.47 +/- 8.98
Iteration 446/500: rewards 18.84 +/- 9.12
Iteration 447/500: rewards 21.41 +/- 10.18
Iteration 448/500: rewards 21.97 +/- 13.89
Iteration 449/500: rewards 18.47 +/- 8.28
Iteration 450/500: rewards 20.47 +/- 9.04
Iteration 451/500: rewards 18.03 +/- 8.78
Iteration 452/500: rewards 20.78 +/- 9.36
Iteration 453/500: rewards 17.19 +/- 7.9
Iteration 454/500: rewards 17.12 +/- 7.74
```

```
Iteration 455/500: rewards 20.25 +/- 9.26
Iteration 456/500: rewards 15.53 +/- 6.66
Iteration 457/500: rewards 20.12 +/- 7.03
Iteration 458/500: rewards 21.47 +/- 9.98
Iteration 459/500: rewards 19.91 +/- 9.14
Iteration 460/500: rewards 20.81 +/- 11.02
Iteration 461/500: rewards 18.06 +/- 7.34
Iteration 462/500: rewards 19.38 +/- 8.1
Iteration 463/500: rewards 18.72 +/- 8.13
Iteration 464/500: rewards 19.28 +/- 9.59
Iteration 465/500: rewards 16.03 +/- 4.74
Iteration 466/500: rewards 19.81 +/- 9.75
Iteration 467/500: rewards 18.19 +/- 5.39
Iteration 468/500: rewards 16.72 +/- 5.38
Iteration 469/500: rewards 17.28 +/- 9.3
Iteration 470/500: rewards 18.97 +/- 7.99
Iteration 471/500: rewards 18.06 +/- 7.47
Iteration 472/500: rewards 19.53 +/- 7.34
Iteration 473/500: rewards 18.62 +/- 9.28
Iteration 474/500: rewards 22.53 +/- 15.77
Iteration 475/500: rewards 18.59 +/- 8.16
Iteration 476/500: rewards 17.53 +/- 7.12
Iteration 477/500: rewards 15.53 +/- 5.81
Iteration 478/500: rewards 19.53 +/- 8.63
Iteration 479/500: rewards 18.88 +/- 11.08
Iteration 480/500: rewards 15.66 +/- 5.22
Iteration 481/500: rewards 19.28 +/- 12.23
Iteration 482/500: rewards 19.47 +/- 8.34
Iteration 483/500: rewards 20.72 +/- 10.01
Iteration 484/500: rewards 21.41 +/- 13.51
Iteration 485/500: rewards 19.81 +/- 8.56
Iteration 486/500: rewards 20.28 +/- 8.51
Iteration 487/500: rewards 18.41 +/- 8.23
Iteration 488/500: rewards 20.5 +/- 8.57
Iteration 489/500: rewards 18.66 +/- 9.3
Iteration 490/500: rewards 18.22 +/- 8.71
Iteration 491/500: rewards 21.81 +/- 13.78
Iteration 492/500: rewards 20.97 +/- 11.57
Iteration 493/500: rewards 19.38 +/- 8.07
Iteration 494/500: rewards 18.91 +/- 8.27
Iteration 495/500: rewards 20.66 +/- 10.17
Iteration 496/500: rewards 20.59 +/- 10.84
Iteration 497/500: rewards 17.97 +/- 6.7
Iteration 498/500: rewards 20.81 +/- 11.44
Iteration 499/500: rewards 17.47 +/- 7.19
Iteration 500/500: rewards 19.25 +/- 8.5
The average reward is 19.05375
the device is: cpu
```

```
------------------------------------------------
The gamma chosen is: 0.95
The value lr chosen is 0.01
The policy lr chosen is 0.01
Iteration 1/500: rewards 19.0 +/- 7.26
Iteration 2/500: rewards 18.16 +/- 9.01
Iteration 3/500: rewards 16.72 +/- 9.7
Iteration 4/500: rewards 18.66 +/- 9.86
Iteration 5/500: rewards 15.94 +/- 7.22
Iteration 6/500: rewards 15.81 +/- 5.49
Iteration 7/500: rewards 14.62 +/- 6.96
Iteration 8/500: rewards 16.66 +/- 8.4
Iteration 9/500: rewards 18.06 +/- 8.77
Iteration 10/500: rewards 16.84 +/- 7.48
Iteration 11/500: rewards 16.16 +/- 8.08
Iteration 12/500: rewards 14.22 +/- 4.07
Iteration 13/500: rewards 15.22 +/- 7.01
Iteration 14/500: rewards 14.97 +/- 4.53
Iteration 15/500: rewards 14.34 +/- 5.3
Iteration 16/500: rewards 15.81 +/- 6.21
Iteration 17/500: rewards 12.88 +/- 3.71
Iteration 18/500: rewards 15.22 +/- 5.64
Iteration 19/500: rewards 13.88 +/- 4.31
Iteration 20/500: rewards 12.47 +/- 3.57
Iteration 21/500: rewards 12.41 +/- 3.08
Iteration 22/500: rewards 11.59 +/- 3.45
Iteration 23/500: rewards 12.09 +/- 2.87
Iteration 24/500: rewards 11.69 +/- 2.64
Iteration 25/500: rewards 11.62 +/- 2.97
Iteration 26/500: rewards 11.69 +/- 3.11
Iteration 27/500: rewards 11.44 +/- 2.97
Iteration 28/500: rewards 11.44 +/- 3.16
Iteration 29/500: rewards 11.19 +/- 2.62
Iteration 30/500: rewards 12.5 +/- 3.71
Iteration 31/500: rewards 11.41 +/- 2.26
Iteration 32/500: rewards 10.97 +/- 2.54
Iteration 33/500: rewards 11.28 +/- 2.86
Iteration 34/500: rewards 11.25 +/- 3.69
Iteration 35/500: rewards 11.88 +/- 3.14
Iteration 36/500: rewards 11.44 +/- 2.75
Iteration 37/500: rewards 12.69 +/- 3.22
Iteration 38/500: rewards 12.38 +/- 2.25
Iteration 39/500: rewards 15.56 +/- 5.61
Iteration 40/500: rewards 14.16 +/- 3.69
Iteration 41/500: rewards 14.94 +/- 8.59
Iteration 42/500: rewards 14.84 +/- 5.72
Iteration 43/500: rewards 14.88 +/- 3.68
Iteration 44/500: rewards 14.94 +/- 4.99
```

```
Iteration 45/500: rewards 15.72 +/- 8.21
Iteration 46/500: rewards 16.12 +/- 6.66
Iteration 47/500: rewards 17.75 +/- 7.07
Iteration 48/500: rewards 19.03 +/- 8.27
Iteration 49/500: rewards 17.19 +/- 7.09
Iteration 50/500: rewards 19.34 +/- 9.3
Iteration 51/500: rewards 19.25 +/- 8.05
Iteration 52/500: rewards 20.25 +/- 12.4
Iteration 53/500: rewards 23.72 +/- 14.02
Iteration 54/500: rewards 24.84 +/- 14.37
Iteration 55/500: rewards 21.81 +/- 13.33
Iteration 56/500: rewards 24.78 +/- 10.9
Iteration 57/500: rewards 28.72 +/- 18.52
Iteration 58/500: rewards 30.84 +/- 15.87
Iteration 59/500: rewards 26.72 +/- 19.32
Iteration 60/500: rewards 26.03 +/- 16.37
Iteration 61/500: rewards 25.28 +/- 12.65
Iteration 62/500: rewards 31.16 +/- 19.64
Iteration 63/500: rewards 23.22 +/- 9.71
Iteration 64/500: rewards 25.91 +/- 10.36
Iteration 65/500: rewards 29.03 +/- 12.11
Iteration 66/500: rewards 27.72 +/- 13.68
Iteration 67/500: rewards 25.56 +/- 13.77
Iteration 68/500: rewards 27.19 +/- 11.0
Iteration 69/500: rewards 35.56 +/- 16.3
Iteration 70/500: rewards 35.62 +/- 15.24
Iteration 71/500: rewards 42.81 +/- 21.07
Iteration 72/500: rewards 41.12 +/- 16.11
Iteration 73/500: rewards 41.22 +/- 22.52
Iteration 74/500: rewards 51.12 +/- 20.43
Iteration 75/500: rewards 50.62 +/- 21.95
Iteration 76/500: rewards 61.34 +/- 25.57
Iteration 77/500: rewards 53.47 +/- 18.45
Iteration 78/500: rewards 59.28 +/- 16.02
Iteration 79/500: rewards 64.16 +/- 15.89
Iteration 80/500: rewards 68.28 +/- 28.43
Iteration 81/500: rewards 73.72 +/- 29.73
Iteration 82/500: rewards 77.38 +/- 35.4
Iteration 83/500: rewards 72.38 +/- 24.87
Iteration 84/500: rewards 74.06 +/- 24.76
Iteration 85/500: rewards 69.38 +/- 26.39
Iteration 86/500: rewards 91.66 +/- 38.05
Iteration 87/500: rewards 96.03 +/- 37.99
Iteration 88/500: rewards 92.03 +/- 36.13
Iteration 89/500: rewards 95.19 +/- 35.5
Iteration 90/500: rewards 89.84 +/- 26.21
Iteration 91/500: rewards 97.19 +/- 40.02
Iteration 92/500: rewards 94.16 +/- 30.03
```

```
Iteration 93/500: rewards 89.28 +/- 33.66
Iteration 94/500: rewards 88.53 +/- 28.51
Iteration 95/500: rewards 84.09 +/- 26.76
Iteration 96/500: rewards 74.91 +/- 18.33
Iteration 97/500: rewards 69.72 +/- 14.53
Iteration 98/500: rewards 64.78 +/- 15.36
Iteration 99/500: rewards 62.0 +/- 14.76
Iteration 100/500: rewards 54.41 +/- 11.79
Iteration 101/500: rewards 52.12 +/- 11.06
Iteration 102/500: rewards 49.56 +/- 10.37
Iteration 103/500: rewards 47.34 +/- 10.12
Iteration 104/500: rewards 44.16 +/- 8.62
Iteration 105/500: rewards 44.66 +/- 7.91
Iteration 106/500: rewards 47.09 +/- 9.15
Iteration 107/500: rewards 45.78 +/- 8.93
Iteration 108/500: rewards 44.75 +/- 10.24
Iteration 109/500: rewards 43.59 +/- 7.47
Iteration 110/500: rewards 43.5 +/- 8.25
Iteration 111/500: rewards 42.69 +/- 7.17
Iteration 112/500: rewards 39.62 +/- 7.33
Iteration 113/500: rewards 35.97 +/- 4.71
Iteration 114/500: rewards 34.53 +/- 4.65
Iteration 115/500: rewards 32.69 +/- 4.43
Iteration 116/500: rewards 29.66 +/- 4.22
Iteration 117/500: rewards 31.53 +/- 4.45
Iteration 118/500: rewards 31.22 +/- 3.92
Iteration 119/500: rewards 30.72 +/- 3.4
Iteration 120/500: rewards 31.12 +/- 5.01
Iteration 121/500: rewards 32.62 +/- 3.66
Iteration 122/500: rewards 34.72 +/- 5.45
Iteration 123/500: rewards 36.19 +/- 5.86
Iteration 124/500: rewards 34.16 +/- 4.32
Iteration 125/500: rewards 35.25 +/- 5.15
Iteration 126/500: rewards 32.69 +/- 4.15
Iteration 127/500: rewards 31.44 +/- 4.31
Iteration 128/500: rewards 30.12 +/- 4.52
Iteration 129/500: rewards 27.84 +/- 3.93
Iteration 130/500: rewards 26.84 +/- 4.98
Iteration 131/500: rewards 25.22 +/- 3.93
Iteration 132/500: rewards 26.12 +/- 4.18
Iteration 133/500: rewards 26.19 +/- 3.97
Iteration 134/500: rewards 28.0 +/- 4.24
Iteration 135/500: rewards 28.56 +/- 4.72
Iteration 136/500: rewards 28.91 +/- 3.84
Iteration 137/500: rewards 30.84 +/- 3.97
Iteration 138/500: rewards 33.41 +/- 5.9
Iteration 139/500: rewards 33.38 +/- 4.87
Iteration 140/500: rewards 31.59 +/- 4.17
```

```
Iteration 141/500: rewards 32.09 +/- 4.83
Iteration 142/500: rewards 32.44 +/- 4.18
Iteration 143/500: rewards 28.94 +/- 4.96
Iteration 144/500: rewards 29.12 +/- 4.57
Iteration 145/500: rewards 26.53 +/- 3.46
Iteration 146/500: rewards 27.0 +/- 4.28
Iteration 147/500: rewards 25.16 +/- 3.96
Iteration 148/500: rewards 26.34 +/- 4.28
Iteration 149/500: rewards 27.47 +/- 4.62
Iteration 150/500: rewards 25.97 +/- 3.69
Iteration 151/500: rewards 27.53 +/- 3.95
Iteration 152/500: rewards 26.31 +/- 3.33
Iteration 153/500: rewards 29.38 +/- 3.83
Iteration 154/500: rewards 30.41 +/- 4.35
Iteration 155/500: rewards 30.44 +/- 3.92
Iteration 156/500: rewards 32.31 +/- 4.65
Iteration 157/500: rewards 32.12 +/- 4.03
Iteration 158/500: rewards 30.91 +/- 4.33
Iteration 159/500: rewards 29.59 +/- 5.21
Iteration 160/500: rewards 29.41 +/- 4.23
Iteration 161/500: rewards 28.09 +/- 4.03
Iteration 162/500: rewards 28.34 +/- 3.52
Iteration 163/500: rewards 27.88 +/- 3.71
Iteration 164/500: rewards 27.19 +/- 4.38
Iteration 165/500: rewards 25.22 +/- 3.71
Iteration 166/500: rewards 26.06 +/- 3.75
Iteration 167/500: rewards 26.41 +/- 3.7
Iteration 168/500: rewards 28.16 +/- 4.4
Iteration 169/500: rewards 28.03 +/- 4.28
Iteration 170/500: rewards 29.16 +/- 3.77
Iteration 171/500: rewards 30.94 +/- 5.08
Iteration 172/500: rewards 32.53 +/- 4.83
Iteration 173/500: rewards 32.62 +/- 4.92
Iteration 174/500: rewards 33.38 +/- 4.53
Iteration 175/500: rewards 32.53 +/- 4.76
Iteration 176/500: rewards 30.5 +/- 5.14
Iteration 177/500: rewards 32.0 +/- 4.12
Iteration 178/500: rewards 31.03 +/- 4.02
Iteration 179/500: rewards 29.56 +/- 4.35
Iteration 180/500: rewards 29.53 +/- 4.88
Iteration 181/500: rewards 28.16 +/- 4.21
Iteration 182/500: rewards 29.62 +/- 3.92
Iteration 183/500: rewards 28.09 +/- 3.32
Iteration 184/500: rewards 29.06 +/- 4.12
Iteration 185/500: rewards 28.25 +/- 4.02
Iteration 186/500: rewards 30.88 +/- 4.52
Iteration 187/500: rewards 30.75 +/- 4.39
Iteration 188/500: rewards 31.09 +/- 4.42
```

```
Iteration 189/500: rewards 32.38 +/- 4.83
Iteration 190/500: rewards 34.91 +/- 5.93
Iteration 191/500: rewards 35.69 +/- 6.72
Iteration 192/500: rewards 37.06 +/- 7.11
Iteration 193/500: rewards 34.75 +/- 6.25
Iteration 194/500: rewards 37.78 +/- 6.09
Iteration 195/500: rewards 34.84 +/- 5.47
Iteration 196/500: rewards 36.56 +/- 5.48
Iteration 197/500: rewards 34.66 +/- 5.76
Iteration 198/500: rewards 34.69 +/- 5.47
Iteration 199/500: rewards 34.06 +/- 5.29
Iteration 200/500: rewards 34.09 +/- 5.79
Iteration 201/500: rewards 33.28 +/- 5.08
Iteration 202/500: rewards 35.19 +/- 5.01
Iteration 203/500: rewards 34.38 +/- 4.91
Iteration 204/500: rewards 37.19 +/- 6.04
Iteration 205/500: rewards 37.0 +/- 6.87
Iteration 206/500: rewards 37.91 +/- 7.41
Iteration 207/500: rewards 41.22 +/- 7.51
Iteration 208/500: rewards 39.38 +/- 6.98
Iteration 209/500: rewards 40.19 +/- 7.76
Iteration 210/500: rewards 40.66 +/- 9.47
Iteration 211/500: rewards 42.91 +/- 8.73
Iteration 212/500: rewards 42.81 +/- 6.7
Iteration 213/500: rewards 42.81 +/- 9.05
Iteration 214/500: rewards 42.91 +/- 8.63
Iteration 215/500: rewards 40.56 +/- 9.16
Iteration 216/500: rewards 37.41 +/- 7.0
Iteration 217/500: rewards 38.47 +/- 6.69
Iteration 218/500: rewards 43.0 +/- 8.94
Iteration 219/500: rewards 36.84 +/- 9.07
Iteration 220/500: rewards 39.44 +/- 7.66
Iteration 221/500: rewards 38.78 +/- 8.46
Iteration 222/500: rewards 38.72 +/- 8.3
Iteration 223/500: rewards 41.16 +/- 9.98
Iteration 224/500: rewards 44.0 +/- 8.81
Iteration 225/500: rewards 40.12 +/- 9.49
Iteration 226/500: rewards 42.56 +/- 10.62
Iteration 227/500: rewards 43.62 +/- 9.37
Iteration 228/500: rewards 45.69 +/- 10.81
Iteration 229/500: rewards 44.44 +/- 10.58
Iteration 230/500: rewards 46.69 +/- 12.08
Iteration 231/500: rewards 47.06 +/- 11.43
Iteration 232/500: rewards 42.56 +/- 8.95
Iteration 233/500: rewards 48.12 +/- 10.91
Iteration 234/500: rewards 43.56 +/- 10.63
Iteration 235/500: rewards 43.41 +/- 9.4
Iteration 236/500: rewards 45.62 +/- 10.92
```

```
Iteration 237/500: rewards 43.12 +/- 8.0
Iteration 238/500: rewards 43.25 +/- 11.85
Iteration 239/500: rewards 42.81 +/- 9.87
Iteration 240/500: rewards 42.22 +/- 10.26
Iteration 241/500: rewards 44.97 +/- 11.1
Iteration 242/500: rewards 45.16 +/- 10.44
Iteration 243/500: rewards 49.56 +/- 13.42
Iteration 244/500: rewards 46.25 +/- 10.89
Iteration 245/500: rewards 43.12 +/- 12.8
Iteration 246/500: rewards 44.12 +/- 10.57
Iteration 247/500: rewards 45.59 +/- 11.04
Iteration 248/500: rewards 49.25 +/- 12.53
Iteration 249/500: rewards 47.34 +/- 12.21
Iteration 250/500: rewards 46.22 +/- 13.11
Iteration 251/500: rewards 47.09 +/- 9.89
Iteration 252/500: rewards 45.34 +/- 14.88
Iteration 253/500: rewards 44.91 +/- 12.84
Iteration 254/500: rewards 45.09 +/- 13.46
Iteration 255/500: rewards 45.91 +/- 10.63
Iteration 256/500: rewards 43.84 +/- 11.37
Iteration 257/500: rewards 46.38 +/- 11.79
Iteration 258/500: rewards 46.69 +/- 12.16
Iteration 259/500: rewards 46.81 +/- 13.36
Iteration 260/500: rewards 49.62 +/- 11.84
Iteration 261/500: rewards 48.81 +/- 13.86
Iteration 262/500: rewards 49.0 +/- 13.14
Iteration 263/500: rewards 48.22 +/- 12.55
Iteration 264/500: rewards 43.38 +/- 10.42
Iteration 265/500: rewards 48.34 +/- 13.15
Iteration 266/500: rewards 50.09 +/- 12.12
Iteration 267/500: rewards 47.03 +/- 11.86
Iteration 268/500: rewards 44.22 +/- 12.16
Iteration 269/500: rewards 47.97 +/- 10.25
Iteration 270/500: rewards 50.66 +/- 14.27
Iteration 271/500: rewards 46.78 +/- 14.22
Iteration 272/500: rewards 52.12 +/- 15.34
Iteration 273/500: rewards 50.53 +/- 11.31
Iteration 274/500: rewards 50.91 +/- 13.99
Iteration 275/500: rewards 48.69 +/- 13.95
Iteration 276/500: rewards 47.88 +/- 12.67
Iteration 277/500: rewards 53.28 +/- 16.83
Iteration 278/500: rewards 48.91 +/- 16.51
Iteration 279/500: rewards 57.06 +/- 21.04
Iteration 280/500: rewards 51.47 +/- 15.78
Iteration 281/500: rewards 49.88 +/- 14.48
Iteration 282/500: rewards 47.03 +/- 10.24
Iteration 283/500: rewards 50.34 +/- 17.92
Iteration 284/500: rewards 48.94 +/- 13.89
```

```
Iteration 285/500: rewards 50.56 +/- 15.91
Iteration 286/500: rewards 49.72 +/- 15.84
Iteration 287/500: rewards 52.19 +/- 14.9
Iteration 288/500: rewards 57.09 +/- 15.66
Iteration 289/500: rewards 49.91 +/- 11.43
Iteration 290/500: rewards 52.28 +/- 14.51
Iteration 291/500: rewards 55.09 +/- 17.0
Iteration 292/500: rewards 54.88 +/- 18.35
Iteration 293/500: rewards 53.72 +/- 18.5
Iteration 294/500: rewards 55.72 +/- 23.94
Iteration 295/500: rewards 57.03 +/- 21.22
Iteration 296/500: rewards 66.0 +/- 24.43
Iteration 297/500: rewards 57.41 +/- 18.08
Iteration 298/500: rewards 66.22 +/- 25.94
Iteration 299/500: rewards 56.97 +/- 20.79
Iteration 300/500: rewards 57.19 +/- 16.41
Iteration 301/500: rewards 68.75 +/- 32.3
Iteration 302/500: rewards 67.34 +/- 28.29
Iteration 303/500: rewards 60.19 +/- 23.11
Iteration 304/500: rewards 63.53 +/- 27.88
Iteration 305/500: rewards 61.5 +/- 21.08
Iteration 306/500: rewards 74.97 +/- 49.23
Iteration 307/500: rewards 81.75 +/- 53.67
Iteration 308/500: rewards 59.22 +/- 17.6
Iteration 309/500: rewards 71.66 +/- 42.96
Iteration 310/500: rewards 71.47 +/- 49.35
Iteration 311/500: rewards 68.5 +/- 33.25
Iteration 312/500: rewards 63.06 +/- 19.52
Iteration 313/500: rewards 74.75 +/- 52.85
Iteration 314/500: rewards 65.53 +/- 27.49
Iteration 315/500: rewards 87.56 +/- 69.56
Iteration 316/500: rewards 70.28 +/- 27.78
Iteration 317/500: rewards 74.5 +/- 31.46
Iteration 318/500: rewards 78.03 +/- 42.94
Iteration 319/500: rewards 80.75 +/- 64.17
Iteration 320/500: rewards 74.0 +/- 35.55
Iteration 321/500: rewards 80.66 +/- 61.4
Iteration 322/500: rewards 73.72 +/- 50.03
Iteration 323/500: rewards 82.12 +/- 54.07
Iteration 324/500: rewards 77.66 +/- 30.38
Iteration 325/500: rewards 108.28 +/- 77.59
Iteration 326/500: rewards 94.31 +/- 85.2
Iteration 327/500: rewards 84.84 +/- 40.3
Iteration 328/500: rewards 103.34 +/- 73.26
Iteration 329/500: rewards 89.28 +/- 51.63
Iteration 330/500: rewards 93.0 +/- 63.66
Iteration 331/500: rewards 95.81 +/- 70.68
Iteration 332/500: rewards 76.97 +/- 25.46
```

```
Iteration 333/500: rewards 106.12 +/- 76.01
Iteration 334/500: rewards 96.09 +/- 39.62
Iteration 335/500: rewards 100.5 +/- 47.68
Iteration 336/500: rewards 113.0 +/- 59.31
Iteration 337/500: rewards 113.75 +/- 50.74
Iteration 338/500: rewards 122.84 +/- 79.71
Iteration 339/500: rewards 125.53 +/- 86.93
Iteration 340/500: rewards 122.19 +/- 61.34
Iteration 341/500: rewards 133.09 +/- 62.65
Iteration 342/500: rewards 152.22 +/- 111.59
Iteration 343/500: rewards 157.59 +/- 91.73
Iteration 344/500: rewards 162.5 +/- 125.73
Iteration 345/500: rewards 157.12 +/- 78.58
Iteration 346/500: rewards 166.41 +/- 104.66
Iteration 347/500: rewards 169.38 +/- 93.91
Iteration 348/500: rewards 164.12 +/- 83.86
Iteration 349/500: rewards 190.0 +/- 90.04
Iteration 350/500: rewards 195.31 +/- 101.81
Iteration 351/500: rewards 179.5 +/- 107.69
Iteration 352/500: rewards 202.25 +/- 120.07
Iteration 353/500: rewards 161.5 +/- 90.49
Iteration 354/500: rewards 226.12 +/- 111.55
Iteration 355/500: rewards 226.56 +/- 107.14
Iteration 356/500: rewards 269.53 +/- 134.68
Iteration 357/500: rewards 201.09 +/- 105.43
Iteration 358/500: rewards 207.03 +/- 98.71
Iteration 359/500: rewards 181.16 +/- 112.55
Iteration 360/500: rewards 173.28 +/- 96.05
Iteration 361/500: rewards 146.47 +/- 76.44
Iteration 362/500: rewards 199.34 +/- 113.49
Iteration 363/500: rewards 161.0 +/- 71.61
Iteration 364/500: rewards 159.03 +/- 72.95
Iteration 365/500: rewards 166.53 +/- 97.05
Iteration 366/500: rewards 127.06 +/- 44.95
Iteration 367/500: rewards 169.22 +/- 102.94
Iteration 368/500: rewards 153.59 +/- 75.04
Iteration 369/500: rewards 144.44 +/- 59.71
Iteration 370/500: rewards 164.25 +/- 91.69
Iteration 371/500: rewards 147.28 +/- 70.8
Iteration 372/500: rewards 172.62 +/- 93.26
Iteration 373/500: rewards 195.12 +/- 92.52
Iteration 374/500: rewards 193.16 +/- 92.5
Iteration 375/500: rewards 191.72 +/- 106.36
Iteration 376/500: rewards 179.66 +/- 88.4
Iteration 377/500: rewards 180.62 +/- 86.42
Iteration 378/500: rewards 166.62 +/- 74.02
Iteration 379/500: rewards 213.47 +/- 89.91
Iteration 380/500: rewards 204.59 +/- 113.25
```

```
Iteration 381/500: rewards 201.44 +/- 110.51
Iteration 382/500: rewards 205.59 +/- 97.62
Iteration 383/500: rewards 187.44 +/- 81.49
Iteration 384/500: rewards 182.81 +/- 82.28
Iteration 385/500: rewards 240.72 +/- 131.46
Iteration 386/500: rewards 177.81 +/- 70.63
Iteration 387/500: rewards 204.47 +/- 107.73
Iteration 388/500: rewards 198.41 +/- 72.68
Iteration 389/500: rewards 171.94 +/- 95.38
Iteration 390/500: rewards 187.09 +/- 125.2
Iteration 391/500: rewards 157.66 +/- 62.65
Iteration 392/500: rewards 180.53 +/- 89.34
Iteration 393/500: rewards 204.19 +/- 99.22
Iteration 394/500: rewards 181.25 +/- 106.12
Iteration 395/500: rewards 137.81 +/- 69.31
Iteration 396/500: rewards 159.34 +/- 80.27
Iteration 397/500: rewards 152.62 +/- 65.96
Iteration 398/500: rewards 122.47 +/- 50.64
Iteration 399/500: rewards 146.34 +/- 65.09
Iteration 400/500: rewards 126.38 +/- 68.95
Iteration 401/500: rewards 117.75 +/- 45.59
Iteration 402/500: rewards 122.03 +/- 55.36
Iteration 403/500: rewards 153.5 +/- 72.75
Iteration 404/500: rewards 115.38 +/- 67.6
Iteration 405/500: rewards 138.25 +/- 81.1
Iteration 406/500: rewards 131.25 +/- 85.65
Iteration 407/500: rewards 93.75 +/- 30.93
Iteration 408/500: rewards 123.69 +/- 67.99
Iteration 409/500: rewards 96.34 +/- 42.26
Iteration 410/500: rewards 126.38 +/- 70.55
Iteration 411/500: rewards 128.03 +/- 72.07
Iteration 412/500: rewards 104.25 +/- 41.64
Iteration 413/500: rewards 86.66 +/- 30.91
Iteration 414/500: rewards 84.72 +/- 19.43
Iteration 415/500: rewards 78.66 +/- 18.06
Iteration 416/500: rewards 81.5 +/- 22.92
Iteration 417/500: rewards 77.84 +/- 15.85
Iteration 418/500: rewards 73.94 +/- 12.78
Iteration 419/500: rewards 75.06 +/- 16.23
Iteration 420/500: rewards 68.84 +/- 12.38
Iteration 421/500: rewards 70.66 +/- 17.42
Iteration 422/500: rewards 65.06 +/- 12.47
Iteration 423/500: rewards 68.62 +/- 12.29
Iteration 424/500: rewards 60.91 +/- 12.39
Iteration 425/500: rewards 63.41 +/- 13.38
Iteration 426/500: rewards 62.22 +/- 12.09
Iteration 427/500: rewards 63.97 +/- 12.1
Iteration 428/500: rewards 59.62 +/- 14.06
```

```
Iteration 429/500: rewards 59.03 +/- 13.03
Iteration 430/500: rewards 55.06 +/- 9.97
Iteration 431/500: rewards 52.88 +/- 10.53
Iteration 432/500: rewards 55.12 +/- 9.64
Iteration 433/500: rewards 51.53 +/- 9.92
Iteration 434/500: rewards 53.97 +/- 7.94
Iteration 435/500: rewards 52.59 +/- 8.68
Iteration 436/500: rewards 48.97 +/- 8.51
Iteration 437/500: rewards 51.66 +/- 8.21
Iteration 438/500: rewards 51.25 +/- 8.8
Iteration 439/500: rewards 51.34 +/- 8.42
Iteration 440/500: rewards 50.56 +/- 9.0
Iteration 441/500: rewards 47.75 +/- 9.05
Iteration 442/500: rewards 47.34 +/- 7.76
Iteration 443/500: rewards 48.88 +/- 8.68
Iteration 444/500: rewards 45.5 +/- 7.83
Iteration 445/500: rewards 45.88 +/- 8.33
Iteration 446/500: rewards 45.91 +/- 8.77
Iteration 447/500: rewards 44.88 +/- 7.18
Iteration 448/500: rewards 45.41 +/- 7.0
Iteration 449/500: rewards 46.81 +/- 8.31
Iteration 450/500: rewards 42.88 +/- 6.82
Iteration 451/500: rewards 44.84 +/- 7.04
Iteration 452/500: rewards 45.22 +/- 6.83
Iteration 453/500: rewards 43.38 +/- 7.81
Iteration 454/500: rewards 44.66 +/- 8.11
Iteration 455/500: rewards 41.84 +/- 6.99
Iteration 456/500: rewards 42.69 +/- 6.48
Iteration 457/500: rewards 42.66 +/- 6.15
Iteration 458/500: rewards 42.09 +/- 5.23
Iteration 459/500: rewards 41.91 +/- 6.29
Iteration 460/500: rewards 42.12 +/- 6.66
Iteration 461/500: rewards 41.47 +/- 6.01
Iteration 462/500: rewards 41.56 +/- 4.89
Iteration 463/500: rewards 38.97 +/- 5.56
Iteration 464/500: rewards 40.22 +/- 5.94
Iteration 465/500: rewards 40.56 +/- 6.81
Iteration 466/500: rewards 37.62 +/- 5.13
Iteration 467/500: rewards 40.19 +/- 5.59
Iteration 468/500: rewards 40.09 +/- 5.78
Iteration 469/500: rewards 39.97 +/- 5.67
Iteration 470/500: rewards 40.34 +/- 4.75
Iteration 471/500: rewards 41.19 +/- 5.85
Iteration 472/500: rewards 40.53 +/- 5.67
Iteration 473/500: rewards 38.66 +/- 5.53
Iteration 474/500: rewards 40.03 +/- 4.64
Iteration 475/500: rewards 38.69 +/- 5.96
Iteration 476/500: rewards 38.28 +/- 4.88
```

```
Iteration 477/500: rewards 38.72 +/- 4.93
Iteration 478/500: rewards 38.66 +/- 4.86
Iteration 479/500: rewards 38.38 +/- 4.17
Iteration 480/500: rewards 37.91 +/- 4.98
Iteration 481/500: rewards 38.59 +/- 3.73
Iteration 482/500: rewards 37.44 +/- 4.37
Iteration 483/500: rewards 35.75 +/- 5.42
Iteration 484/500: rewards 37.19 +/- 5.8
Iteration 485/500: rewards 36.16 +/- 4.64
Iteration 486/500: rewards 36.28 +/- 5.96
Iteration 487/500: rewards 37.66 +/- 5.27
Iteration 488/500: rewards 34.91 +/- 5.29
Iteration 489/500: rewards 35.09 +/- 5.43
Iteration 490/500: rewards 34.69 +/- 4.54
Iteration 491/500: rewards 37.12 +/- 5.65
Iteration 492/500: rewards 35.91 +/- 4.47
Iteration 493/500: rewards 36.28 +/- 4.89
Iteration 494/500: rewards 36.66 +/- 4.85
Iteration 495/500: rewards 36.31 +/- 5.1
Iteration 496/500: rewards 37.5 +/- 4.36
Iteration 497/500: rewards 37.44 +/- 5.12
Iteration 498/500: rewards 36.38 +/- 4.94
Iteration 499/500: rewards 37.78 +/- 5.24
Iteration 500/500: rewards 38.31 +/- 5.36
The average reward is 39.1375
the device is: cpu
------------------------------------------------
The gamma chosen is: 0.95
The value lr chosen is 0.001
The policy lr chosen is 0.01
Iteration 1/500: rewards 19.0 +/- 7.26
Iteration 2/500: rewards 18.16 +/- 9.01
Iteration 3/500: rewards 16.72 +/- 9.7
Iteration 4/500: rewards 18.66 +/- 9.86
Iteration 5/500: rewards 15.47 +/- 6.36
Iteration 6/500: rewards 15.81 +/- 5.5
Iteration 7/500: rewards 14.06 +/- 4.53
Iteration 8/500: rewards 17.56 +/- 7.93
Iteration 9/500: rewards 17.78 +/- 6.94
Iteration 10/500: rewards 17.5 +/- 6.2
Iteration 11/500: rewards 18.03 +/- 8.38
Iteration 12/500: rewards 15.88 +/- 6.41
Iteration 13/500: rewards 19.38 +/- 8.77
Iteration 14/500: rewards 19.91 +/- 8.47
Iteration 15/500: rewards 17.44 +/- 9.91
Iteration 16/500: rewards 18.38 +/- 9.4
Iteration 17/500: rewards 18.38 +/- 8.0
Iteration 18/500: rewards 17.91 +/- 9.55
```

```
Iteration 19/500: rewards 17.09 +/- 6.56
Iteration 20/500: rewards 18.31 +/- 8.91
Iteration 21/500: rewards 18.28 +/- 10.08
Iteration 22/500: rewards 18.28 +/- 8.25
Iteration 23/500: rewards 20.09 +/- 11.31
Iteration 24/500: rewards 19.09 +/- 9.23
Iteration 25/500: rewards 16.34 +/- 6.79
Iteration 26/500: rewards 15.41 +/- 6.16
Iteration 27/500: rewards 16.0 +/- 5.08
Iteration 28/500: rewards 19.78 +/- 11.95
Iteration 29/500: rewards 15.81 +/- 5.9
Iteration 30/500: rewards 19.38 +/- 9.45
Iteration 31/500: rewards 18.47 +/- 6.31
Iteration 32/500: rewards 16.09 +/- 6.61
Iteration 33/500: rewards 15.31 +/- 4.33
Iteration 34/500: rewards 16.72 +/- 8.04
Iteration 35/500: rewards 19.09 +/- 10.81
Iteration 36/500: rewards 15.03 +/- 5.24
Iteration 37/500: rewards 14.94 +/- 5.97
Iteration 38/500: rewards 14.84 +/- 4.63
Iteration 39/500: rewards 15.72 +/- 7.24
Iteration 40/500: rewards 17.31 +/- 7.02
Iteration 41/500: rewards 14.47 +/- 6.71
Iteration 42/500: rewards 14.28 +/- 4.08
Iteration 43/500: rewards 14.06 +/- 5.33
Iteration 44/500: rewards 12.69 +/- 3.35
Iteration 45/500: rewards 13.0 +/- 5.11
Iteration 46/500: rewards 12.75 +/- 3.08
Iteration 47/500: rewards 12.47 +/- 4.74
Iteration 48/500: rewards 14.22 +/- 5.34
Iteration 49/500: rewards 11.62 +/- 2.34
Iteration 50/500: rewards 12.66 +/- 3.88
Iteration 51/500: rewards 11.88 +/- 3.35
Iteration 52/500: rewards 11.19 +/- 2.75
Iteration 53/500: rewards 11.41 +/- 2.52
Iteration 54/500: rewards 11.66 +/- 4.16
Iteration 55/500: rewards 11.59 +/- 4.66
Iteration 56/500: rewards 11.22 +/- 3.01
Iteration 57/500: rewards 10.84 +/- 2.2
Iteration 58/500: rewards 10.72 +/- 3.03
Iteration 59/500: rewards 10.91 +/- 2.08
Iteration 60/500: rewards 10.38 +/- 2.04
Iteration 61/500: rewards 10.16 +/- 1.64
Iteration 62/500: rewards 10.44 +/- 2.12
Iteration 63/500: rewards 10.06 +/- 1.54
Iteration 64/500: rewards 10.25 +/- 2.03
Iteration 65/500: rewards 10.06 +/- 1.39
Iteration 66/500: rewards 10.44 +/- 1.89
```

```
Iteration 67/500: rewards 9.56 +/- 1.32
Iteration 68/500: rewards 10.25 +/- 2.02
Iteration 69/500: rewards 10.34 +/- 2.26
Iteration 70/500: rewards 9.78 +/- 1.02
Iteration 71/500: rewards 10.25 +/- 2.21
Iteration 72/500: rewards 9.91 +/- 1.23
Iteration 73/500: rewards 9.75 +/- 1.75
Iteration 74/500: rewards 10.25 +/- 3.02
Iteration 75/500: rewards 9.88 +/- 1.34
Iteration 76/500: rewards 10.12 +/- 1.36
Iteration 77/500: rewards 10.25 +/- 2.14
Iteration 78/500: rewards 9.78 +/- 1.34
Iteration 79/500: rewards 9.84 +/- 1.35
Iteration 80/500: rewards 10.0 +/- 1.27
Iteration 81/500: rewards 9.84 +/- 1.48
Iteration 82/500: rewards 10.31 +/- 2.55
Iteration 83/500: rewards 10.03 +/- 1.38
Iteration 84/500: rewards 9.81 +/- 2.44
Iteration 85/500: rewards 9.81 +/- 1.51
Iteration 86/500: rewards 9.94 +/- 1.43
Iteration 87/500: rewards 10.0 +/- 1.44
Iteration 88/500: rewards 10.06 +/- 1.54
Iteration 89/500: rewards 9.72 +/- 1.12
Iteration 90/500: rewards 10.38 +/- 1.96
Iteration 91/500: rewards 10.03 +/- 1.19
Iteration 92/500: rewards 9.5 +/- 1.39
Iteration 93/500: rewards 9.5 +/- 1.06
Iteration 94/500: rewards 9.59 +/- 1.25
Iteration 95/500: rewards 10.47 +/- 3.0
Iteration 96/500: rewards 10.19 +/- 2.93
Iteration 97/500: rewards 9.5 +/- 0.79
Iteration 98/500: rewards 9.75 +/- 1.22
Iteration 99/500: rewards 9.84 +/- 1.35
Iteration 100/500: rewards 9.94 +/- 1.27
Iteration 101/500: rewards 9.5 +/- 1.0
Iteration 102/500: rewards 9.91 +/- 1.18
Iteration 103/500: rewards 9.56 +/- 1.17
Iteration 104/500: rewards 9.91 +/- 1.1
Iteration 105/500: rewards 10.16 +/- 2.54
Iteration 106/500: rewards 9.56 +/- 0.93
Iteration 107/500: rewards 9.97 +/- 1.45
Iteration 108/500: rewards 9.91 +/- 1.16
Iteration 109/500: rewards 9.38 +/- 0.86
Iteration 110/500: rewards 9.53 +/- 0.9
Iteration 111/500: rewards 9.34 +/- 0.99
Iteration 112/500: rewards 9.47 +/- 1.12
Iteration 113/500: rewards 9.78 +/- 1.54
Iteration 114/500: rewards 9.72 +/- 0.87
```

```
Iteration 115/500: rewards 9.69 +/- 1.18
Iteration 116/500: rewards 9.84 +/- 1.03
Iteration 117/500: rewards 9.38 +/- 0.93
Iteration 118/500: rewards 9.91 +/- 1.59
Iteration 119/500: rewards 9.72 +/- 0.98
Iteration 120/500: rewards 10.12 +/- 1.41
Iteration 121/500: rewards 9.72 +/- 1.28
Iteration 122/500: rewards 9.97 +/- 1.24
Iteration 123/500: rewards 9.78 +/- 1.24
Iteration 124/500: rewards 9.81 +/- 1.4
Iteration 125/500: rewards 9.94 +/- 2.33
Iteration 126/500: rewards 10.0 +/- 1.5
Iteration 127/500: rewards 9.72 +/- 1.26
Iteration 128/500: rewards 9.75 +/- 1.27
Iteration 129/500: rewards 10.0 +/- 0.94
Iteration 130/500: rewards 10.41 +/- 2.32
Iteration 131/500: rewards 10.12 +/- 1.58
Iteration 132/500: rewards 9.84 +/- 1.44
Iteration 133/500: rewards 9.62 +/- 0.96
Iteration 134/500: rewards 9.78 +/- 1.24
Iteration 135/500: rewards 10.0 +/- 1.56
Iteration 136/500: rewards 9.69 +/- 1.21
Iteration 137/500: rewards 9.66 +/- 1.08
Iteration 138/500: rewards 9.88 +/- 1.58
Iteration 139/500: rewards 9.69 +/- 1.07
Iteration 140/500: rewards 9.97 +/- 1.31
Iteration 141/500: rewards 10.0 +/- 2.57
Iteration 142/500: rewards 9.84 +/- 1.6
Iteration 143/500: rewards 10.16 +/- 1.48
Iteration 144/500: rewards 10.06 +/- 1.56
Iteration 145/500: rewards 10.53 +/- 1.85
Iteration 146/500: rewards 10.12 +/- 1.83
Iteration 147/500: rewards 9.81 +/- 1.18
Iteration 148/500: rewards 10.16 +/- 2.05
Iteration 149/500: rewards 9.97 +/- 1.9
Iteration 150/500: rewards 10.56 +/- 1.56
Iteration 151/500: rewards 10.38 +/- 1.65
Iteration 152/500: rewards 11.34 +/- 2.19
Iteration 153/500: rewards 10.59 +/- 2.42
Iteration 154/500: rewards 10.38 +/- 1.8
Iteration 155/500: rewards 10.88 +/- 2.19
Iteration 156/500: rewards 11.0 +/- 3.51
Iteration 157/500: rewards 11.62 +/- 2.98
Iteration 158/500: rewards 11.28 +/- 2.47
Iteration 159/500: rewards 11.28 +/- 2.96
Iteration 160/500: rewards 11.59 +/- 3.38
Iteration 161/500: rewards 11.69 +/- 2.62
Iteration 162/500: rewards 11.75 +/- 4.17
```

```
Iteration 163/500: rewards 12.72 +/- 4.12
Iteration 164/500: rewards 12.53 +/- 4.66
Iteration 165/500: rewards 13.12 +/- 4.56
Iteration 166/500: rewards 14.94 +/- 5.2
Iteration 167/500: rewards 15.66 +/- 7.37
Iteration 168/500: rewards 15.44 +/- 5.99
Iteration 169/500: rewards 15.41 +/- 7.51
Iteration 170/500: rewards 17.16 +/- 6.66
Iteration 171/500: rewards 16.09 +/- 7.72
Iteration 172/500: rewards 15.22 +/- 6.53
Iteration 173/500: rewards 16.88 +/- 6.94
Iteration 174/500: rewards 18.34 +/- 6.67
Iteration 175/500: rewards 18.12 +/- 9.35
Iteration 176/500: rewards 20.81 +/- 10.87
Iteration 177/500: rewards 16.03 +/- 5.42
Iteration 178/500: rewards 18.09 +/- 6.42
Iteration 179/500: rewards 18.19 +/- 5.02
Iteration 180/500: rewards 17.38 +/- 6.51
Iteration 181/500: rewards 19.28 +/- 7.56
Iteration 182/500: rewards 14.81 +/- 4.25
Iteration 183/500: rewards 16.03 +/- 5.02
Iteration 184/500: rewards 16.25 +/- 4.79
Iteration 185/500: rewards 17.03 +/- 5.31
Iteration 186/500: rewards 15.12 +/- 3.72
Iteration 187/500: rewards 15.94 +/- 5.04
Iteration 188/500: rewards 16.25 +/- 4.15
Iteration 189/500: rewards 18.56 +/- 5.17
Iteration 190/500: rewards 17.41 +/- 5.09
Iteration 191/500: rewards 20.31 +/- 7.68
Iteration 192/500: rewards 19.09 +/- 6.09
Iteration 193/500: rewards 19.88 +/- 6.92
Iteration 194/500: rewards 21.5 +/- 7.3
Iteration 195/500: rewards 20.0 +/- 7.46
Iteration 196/500: rewards 21.56 +/- 6.77
Iteration 197/500: rewards 23.5 +/- 10.23
Iteration 198/500: rewards 22.94 +/- 7.19
Iteration 199/500: rewards 24.56 +/- 8.65
Iteration 200/500: rewards 26.94 +/- 10.56
Iteration 201/500: rewards 30.06 +/- 12.9
Iteration 202/500: rewards 30.78 +/- 11.21
Iteration 203/500: rewards 35.75 +/- 21.92
Iteration 204/500: rewards 37.16 +/- 21.48
Iteration 205/500: rewards 38.81 +/- 21.08
Iteration 206/500: rewards 41.91 +/- 17.76
Iteration 207/500: rewards 36.0 +/- 10.93
Iteration 208/500: rewards 44.5 +/- 23.53
Iteration 209/500: rewards 41.94 +/- 15.46
Iteration 210/500: rewards 45.06 +/- 19.75
```

```
Iteration 211/500: rewards 39.38 +/- 16.15
Iteration 212/500: rewards 38.16 +/- 9.53
Iteration 213/500: rewards 40.19 +/- 13.5
Iteration 214/500: rewards 37.75 +/- 15.81
Iteration 215/500: rewards 43.78 +/- 16.85
Iteration 216/500: rewards 47.97 +/- 18.6
Iteration 217/500: rewards 45.78 +/- 22.35
Iteration 218/500: rewards 42.62 +/- 28.53
Iteration 219/500: rewards 57.91 +/- 30.34
Iteration 220/500: rewards 49.5 +/- 21.56
Iteration 221/500: rewards 58.44 +/- 26.35
Iteration 222/500: rewards 51.72 +/- 16.86
Iteration 223/500: rewards 50.66 +/- 13.6
Iteration 224/500: rewards 46.72 +/- 14.45
Iteration 225/500: rewards 56.31 +/- 18.9
Iteration 226/500: rewards 53.09 +/- 15.78
Iteration 227/500: rewards 48.38 +/- 16.35
Iteration 228/500: rewards 50.38 +/- 23.22
Iteration 229/500: rewards 48.41 +/- 18.32
Iteration 230/500: rewards 49.78 +/- 27.68
Iteration 231/500: rewards 43.25 +/- 20.66
Iteration 232/500: rewards 54.19 +/- 37.84
Iteration 233/500: rewards 42.19 +/- 19.71
Iteration 234/500: rewards 52.53 +/- 27.85
Iteration 235/500: rewards 47.44 +/- 22.3
Iteration 236/500: rewards 46.31 +/- 20.32
Iteration 237/500: rewards 46.53 +/- 16.26
Iteration 238/500: rewards 52.88 +/- 26.76
Iteration 239/500: rewards 49.09 +/- 17.23
Iteration 240/500: rewards 60.41 +/- 35.96
Iteration 241/500: rewards 51.78 +/- 20.4
Iteration 242/500: rewards 51.44 +/- 21.16
Iteration 243/500: rewards 49.41 +/- 20.55
Iteration 244/500: rewards 50.31 +/- 20.15
Iteration 245/500: rewards 56.12 +/- 19.59
Iteration 246/500: rewards 51.81 +/- 17.05
Iteration 247/500: rewards 52.25 +/- 19.1
Iteration 248/500: rewards 47.38 +/- 15.99
Iteration 249/500: rewards 45.22 +/- 13.31
Iteration 250/500: rewards 52.16 +/- 22.44
Iteration 251/500: rewards 45.41 +/- 12.24
Iteration 252/500: rewards 52.53 +/- 18.32
Iteration 253/500: rewards 49.88 +/- 14.74
Iteration 254/500: rewards 54.03 +/- 18.22
Iteration 255/500: rewards 47.44 +/- 12.56
Iteration 256/500: rewards 54.44 +/- 18.77
Iteration 257/500: rewards 54.0 +/- 19.29
Iteration 258/500: rewards 51.81 +/- 22.49
```

```
Iteration 259/500: rewards 53.31 +/- 20.84
Iteration 260/500: rewards 47.53 +/- 12.87
Iteration 261/500: rewards 50.84 +/- 16.99
Iteration 262/500: rewards 53.31 +/- 20.52
Iteration 263/500: rewards 51.66 +/- 14.33
Iteration 264/500: rewards 55.19 +/- 12.71
Iteration 265/500: rewards 53.31 +/- 12.96
Iteration 266/500: rewards 57.09 +/- 20.62
Iteration 267/500: rewards 58.66 +/- 19.71
Iteration 268/500: rewards 70.12 +/- 38.4
Iteration 269/500: rewards 62.5 +/- 22.13
Iteration 270/500: rewards 61.12 +/- 24.11
Iteration 271/500: rewards 82.69 +/- 54.09
Iteration 272/500: rewards 71.09 +/- 45.78
Iteration 273/500: rewards 62.03 +/- 18.07
Iteration 274/500: rewards 62.81 +/- 21.86
Iteration 275/500: rewards 84.41 +/- 51.39
Iteration 276/500: rewards 62.28 +/- 19.48
Iteration 277/500: rewards 80.66 +/- 57.29
Iteration 278/500: rewards 65.81 +/- 22.41
Iteration 279/500: rewards 75.53 +/- 45.68
Iteration 280/500: rewards 84.31 +/- 60.28
Iteration 281/500: rewards 82.97 +/- 50.57
Iteration 282/500: rewards 97.09 +/- 59.57
Iteration 283/500: rewards 83.78 +/- 68.52
Iteration 284/500: rewards 89.91 +/- 48.28
Iteration 285/500: rewards 94.06 +/- 45.13
Iteration 286/500: rewards 98.78 +/- 58.58
Iteration 287/500: rewards 95.09 +/- 67.29
Iteration 288/500: rewards 107.75 +/- 78.01
Iteration 289/500: rewards 125.12 +/- 107.6
Iteration 290/500: rewards 116.88 +/- 80.89
Iteration 291/500: rewards 110.53 +/- 71.9
Iteration 292/500: rewards 104.47 +/- 63.94
Iteration 293/500: rewards 131.91 +/- 66.59
Iteration 294/500: rewards 118.53 +/- 61.42
Iteration 295/500: rewards 143.84 +/- 103.62
Iteration 296/500: rewards 126.12 +/- 73.19
Iteration 297/500: rewards 154.03 +/- 127.97
Iteration 298/500: rewards 114.5 +/- 103.06
Iteration 299/500: rewards 116.97 +/- 62.56
Iteration 300/500: rewards 126.81 +/- 60.55
Iteration 301/500: rewards 114.38 +/- 95.41
Iteration 302/500: rewards 99.38 +/- 47.07
Iteration 303/500: rewards 120.75 +/- 93.34
Iteration 304/500: rewards 131.34 +/- 94.12
Iteration 305/500: rewards 128.88 +/- 96.73
Iteration 306/500: rewards 139.97 +/- 84.03
```

```
Iteration 307/500: rewards 108.78 +/- 65.63
Iteration 308/500: rewards 122.97 +/- 88.02
Iteration 309/500: rewards 118.0 +/- 83.45
Iteration 310/500: rewards 134.53 +/- 94.52
Iteration 311/500: rewards 84.72 +/- 46.2
Iteration 312/500: rewards 78.44 +/- 31.21
Iteration 313/500: rewards 90.28 +/- 70.26
Iteration 314/500: rewards 67.03 +/- 16.88
Iteration 315/500: rewards 86.12 +/- 71.24
Iteration 316/500: rewards 70.31 +/- 22.91
Iteration 317/500: rewards 71.69 +/- 23.58
Iteration 318/500: rewards 73.28 +/- 27.62
Iteration 319/500: rewards 71.66 +/- 30.41
Iteration 320/500: rewards 90.62 +/- 89.17
Iteration 321/500: rewards 65.25 +/- 26.06
Iteration 322/500: rewards 69.72 +/- 27.48
Iteration 323/500: rewards 82.09 +/- 79.17
Iteration 324/500: rewards 71.06 +/- 20.06
Iteration 325/500: rewards 75.91 +/- 27.16
Iteration 326/500: rewards 96.44 +/- 74.3
Iteration 327/500: rewards 119.88 +/- 106.19
Iteration 328/500: rewards 95.69 +/- 60.66
Iteration 329/500: rewards 92.12 +/- 84.51
Iteration 330/500: rewards 77.91 +/- 35.45
Iteration 331/500: rewards 86.28 +/- 43.58
Iteration 332/500: rewards 97.81 +/- 97.17
Iteration 333/500: rewards 90.16 +/- 77.33
Iteration 334/500: rewards 117.94 +/- 88.4
Iteration 335/500: rewards 97.72 +/- 63.55
Iteration 336/500: rewards 122.56 +/- 98.05
Iteration 337/500: rewards 143.0 +/- 120.77
Iteration 338/500: rewards 118.78 +/- 74.78
Iteration 339/500: rewards 103.72 +/- 74.75
Iteration 340/500: rewards 110.31 +/- 81.3
Iteration 341/500: rewards 117.72 +/- 82.35
Iteration 342/500: rewards 123.25 +/- 106.64
Iteration 343/500: rewards 131.94 +/- 110.86
Iteration 344/500: rewards 101.41 +/- 67.72
Iteration 345/500: rewards 138.78 +/- 125.61
Iteration 346/500: rewards 114.47 +/- 63.77
Iteration 347/500: rewards 128.84 +/- 94.99
Iteration 348/500: rewards 103.56 +/- 60.23
Iteration 349/500: rewards 120.44 +/- 77.84
Iteration 350/500: rewards 136.56 +/- 100.32
Iteration 351/500: rewards 134.22 +/- 100.86
Iteration 352/500: rewards 164.38 +/- 121.53
Iteration 353/500: rewards 104.94 +/- 81.99
Iteration 354/500: rewards 145.44 +/- 116.75
```

```
Iteration 355/500: rewards 122.5 +/- 59.22
Iteration 356/500: rewards 158.03 +/- 118.03
Iteration 357/500: rewards 137.66 +/- 116.47
Iteration 358/500: rewards 170.75 +/- 113.11
Iteration 359/500: rewards 152.75 +/- 114.77
Iteration 360/500: rewards 157.25 +/- 127.45
Iteration 361/500: rewards 143.97 +/- 123.48
Iteration 362/500: rewards 155.44 +/- 116.81
Iteration 363/500: rewards 132.59 +/- 96.64
Iteration 364/500: rewards 143.91 +/- 102.01
Iteration 365/500: rewards 136.22 +/- 103.16
Iteration 366/500: rewards 129.75 +/- 93.8
Iteration 367/500: rewards 176.66 +/- 139.12
Iteration 368/500: rewards 168.16 +/- 148.87
Iteration 369/500: rewards 189.16 +/- 145.95
Iteration 370/500: rewards 144.16 +/- 125.64
Iteration 371/500: rewards 136.09 +/- 99.07
Iteration 372/500: rewards 139.06 +/- 95.84
Iteration 373/500: rewards 121.25 +/- 75.9
Iteration 374/500: rewards 153.88 +/- 114.59
Iteration 375/500: rewards 123.03 +/- 86.06
Iteration 376/500: rewards 113.69 +/- 66.75
Iteration 377/500: rewards 121.03 +/- 99.42
Iteration 378/500: rewards 98.09 +/- 49.28
Iteration 379/500: rewards 128.38 +/- 100.74
Iteration 380/500: rewards 95.0 +/- 29.79
Iteration 381/500: rewards 97.22 +/- 39.24
Iteration 382/500: rewards 102.69 +/- 49.79
Iteration 383/500: rewards 119.28 +/- 89.15
Iteration 384/500: rewards 99.0 +/- 57.35
Iteration 385/500: rewards 110.25 +/- 80.22
Iteration 386/500: rewards 106.0 +/- 55.34
Iteration 387/500: rewards 109.09 +/- 50.72
Iteration 388/500: rewards 125.41 +/- 82.12
Iteration 389/500: rewards 114.25 +/- 81.07
Iteration 390/500: rewards 127.41 +/- 76.68
Iteration 391/500: rewards 105.47 +/- 34.92
Iteration 392/500: rewards 117.56 +/- 78.33
Iteration 393/500: rewards 135.44 +/- 85.71
Iteration 394/500: rewards 140.12 +/- 71.95
Iteration 395/500: rewards 133.78 +/- 98.96
Iteration 396/500: rewards 177.06 +/- 81.8
Iteration 397/500: rewards 193.91 +/- 131.47
Iteration 398/500: rewards 127.62 +/- 80.54
Iteration 399/500: rewards 143.53 +/- 94.12
Iteration 400/500: rewards 186.69 +/- 139.74
Iteration 401/500: rewards 148.19 +/- 88.75
Iteration 402/500: rewards 127.84 +/- 62.41
```

```
Iteration 403/500: rewards 126.78 +/- 45.9
Iteration 404/500: rewards 144.53 +/- 107.57
Iteration 405/500: rewards 140.66 +/- 45.21
Iteration 406/500: rewards 137.09 +/- 84.78
Iteration 407/500: rewards 140.03 +/- 85.26
Iteration 408/500: rewards 133.91 +/- 53.45
Iteration 409/500: rewards 125.97 +/- 52.24
Iteration 410/500: rewards 131.47 +/- 37.66
Iteration 411/500: rewards 134.56 +/- 30.15
Iteration 412/500: rewards 156.44 +/- 38.01
Iteration 413/500: rewards 163.84 +/- 73.88
Iteration 414/500: rewards 158.62 +/- 27.66
Iteration 415/500: rewards 148.81 +/- 19.21
Iteration 416/500: rewards 157.38 +/- 25.73
Iteration 417/500: rewards 177.97 +/- 54.15
Iteration 418/500: rewards 180.03 +/- 37.21
Iteration 419/500: rewards 192.41 +/- 65.6
Iteration 420/500: rewards 163.16 +/- 20.42
Iteration 421/500: rewards 183.41 +/- 65.78
Iteration 422/500: rewards 158.38 +/- 26.44
Iteration 423/500: rewards 162.69 +/- 21.49
Iteration 424/500: rewards 153.16 +/- 19.03
Iteration 425/500: rewards 154.16 +/- 16.15
Iteration 426/500: rewards 150.88 +/- 18.36
Iteration 427/500: rewards 164.72 +/- 30.84
Iteration 428/500: rewards 183.75 +/- 40.74
Iteration 429/500: rewards 200.25 +/- 39.73
Iteration 430/500: rewards 233.16 +/- 22.37
Iteration 431/500: rewards 233.59 +/- 21.43
Iteration 432/500: rewards 236.19 +/- 21.65
Iteration 433/500: rewards 226.19 +/- 24.46
Iteration 434/500: rewards 235.09 +/- 14.04
Iteration 435/500: rewards 224.75 +/- 21.89
Iteration 436/500: rewards 210.09 +/- 11.85
Iteration 437/500: rewards 191.34 +/- 24.59
Iteration 438/500: rewards 183.28 +/- 32.88
Iteration 439/500: rewards 182.66 +/- 35.66
Iteration 440/500: rewards 151.16 +/- 33.56
Iteration 441/500: rewards 164.56 +/- 50.42
Iteration 442/500: rewards 146.66 +/- 42.17
Iteration 443/500: rewards 134.38 +/- 47.27
Iteration 444/500: rewards 122.06 +/- 22.87
Iteration 445/500: rewards 130.06 +/- 39.99
Iteration 446/500: rewards 118.72 +/- 20.79
Iteration 447/500: rewards 120.22 +/- 15.66
Iteration 448/500: rewards 120.16 +/- 17.45
Iteration 449/500: rewards 123.12 +/- 19.86
Iteration 450/500: rewards 120.41 +/- 29.47
```

```
Iteration 451/500: rewards 119.62 +/- 18.0
Iteration 452/500: rewards 132.31 +/- 45.71
Iteration 453/500: rewards 122.38 +/- 29.91
Iteration 454/500: rewards 119.78 +/- 21.73
Iteration 455/500: rewards 118.47 +/- 29.04
Iteration 456/500: rewards 111.0 +/- 14.67
Iteration 457/500: rewards 113.78 +/- 18.58
Iteration 458/500: rewards 109.34 +/- 12.7
Iteration 459/500: rewards 108.78 +/- 13.63
Iteration 460/500: rewards 110.59 +/- 14.72
Iteration 461/500: rewards 111.5 +/- 15.64
Iteration 462/500: rewards 112.25 +/- 16.9
Iteration 463/500: rewards 106.06 +/- 12.25
Iteration 464/500: rewards 107.69 +/- 11.09
Iteration 465/500: rewards 112.84 +/- 18.59
Iteration 466/500: rewards 106.53 +/- 13.38
Iteration 467/500: rewards 110.12 +/- 13.72
Iteration 468/500: rewards 109.97 +/- 15.9
Iteration 469/500: rewards 108.72 +/- 17.85
Iteration 470/500: rewards 111.78 +/- 16.51
Iteration 471/500: rewards 102.19 +/- 24.59
Iteration 472/500: rewards 88.47 +/- 25.09
Iteration 473/500: rewards 80.75 +/- 22.96
Iteration 474/500: rewards 73.66 +/- 23.25
Iteration 475/500: rewards 65.34 +/- 13.76
Iteration 476/500: rewards 59.91 +/- 13.37
Iteration 477/500: rewards 54.0 +/- 10.53
Iteration 478/500: rewards 55.81 +/- 9.19
Iteration 479/500: rewards 56.0 +/- 10.0
Iteration 480/500: rewards 52.28 +/- 8.83
Iteration 481/500: rewards 55.03 +/- 8.38
Iteration 482/500: rewards 52.31 +/- 9.39
Iteration 483/500: rewards 51.84 +/- 9.96
Iteration 484/500: rewards 53.06 +/- 10.7
Iteration 485/500: rewards 51.75 +/- 9.79
Iteration 486/500: rewards 52.34 +/- 9.52
Iteration 487/500: rewards 53.59 +/- 9.79
Iteration 488/500: rewards 50.31 +/- 10.08
Iteration 489/500: rewards 51.41 +/- 10.19
Iteration 490/500: rewards 52.22 +/- 10.23
Iteration 491/500: rewards 55.75 +/- 10.49
Iteration 492/500: rewards 54.59 +/- 11.36
Iteration 493/500: rewards 55.94 +/- 10.41
Iteration 494/500: rewards 57.09 +/- 11.49
Iteration 495/500: rewards 55.97 +/- 9.3
Iteration 496/500: rewards 60.38 +/- 11.75
Iteration 497/500: rewards 61.62 +/- 12.3
Iteration 498/500: rewards 63.19 +/- 13.99
```

```
Iteration 499/500: rewards 66.97 +/- 11.82
Iteration 500/500: rewards 69.78 +/- 15.33
The average reward is 81.541875
the device is: cpu
--------------------------------------------------
The gamma chosen is: 0.95
The value lr chosen is 0.0001
The policy lr chosen is 0.01
Iteration 1/500: rewards 19.0 +/- 7.26
Iteration 2/500: rewards 18.16 +/- 9.01
Iteration 3/500: rewards 16.72 +/- 9.7
Iteration 4/500: rewards 18.66 +/- 9.86
Iteration 5/500: rewards 15.47 +/- 6.36
Iteration 6/500: rewards 15.81 +/- 5.5
Iteration 7/500: rewards 14.06 +/- 4.53
Iteration 8/500: rewards 17.56 +/- 7.93
Iteration 9/500: rewards 17.66 +/- 7.17
Iteration 10/500: rewards 17.62 +/- 6.2
Iteration 11/500: rewards 17.84 +/- 8.31
Iteration 12/500: rewards 17.16 +/- 7.04
Iteration 13/500: rewards 20.47 +/- 9.32
Iteration 14/500: rewards 20.56 +/- 12.04
Iteration 15/500: rewards 18.38 +/- 12.81
Iteration 16/500: rewards 16.53 +/- 4.86
Iteration 17/500: rewards 19.88 +/- 9.71
Iteration 18/500: rewards 17.97 +/- 6.54
Iteration 19/500: rewards 17.09 +/- 5.06
Iteration 20/500: rewards 16.5 +/- 7.2
Iteration 21/500: rewards 16.59 +/- 6.42
Iteration 22/500: rewards 19.06 +/- 9.31
Iteration 23/500: rewards 18.34 +/- 9.01
Iteration 24/500: rewards 24.06 +/- 16.79
Iteration 25/500: rewards 18.75 +/- 8.46
Iteration 26/500: rewards 21.28 +/- 11.98
Iteration 27/500: rewards 16.91 +/- 5.5
Iteration 28/500: rewards 20.75 +/- 10.64
Iteration 29/500: rewards 20.69 +/- 9.19
Iteration 30/500: rewards 18.47 +/- 6.45
Iteration 31/500: rewards 19.81 +/- 8.49
Iteration 32/500: rewards 19.56 +/- 10.75
Iteration 33/500: rewards 16.06 +/- 5.92
Iteration 34/500: rewards 17.25 +/- 6.27
Iteration 35/500: rewards 18.03 +/- 6.63
Iteration 36/500: rewards 21.28 +/- 9.61
Iteration 37/500: rewards 21.12 +/- 12.84
Iteration 38/500: rewards 17.41 +/- 8.76
Iteration 39/500: rewards 18.94 +/- 7.91
Iteration 40/500: rewards 20.12 +/- 8.12
```

```
Iteration 41/500: rewards 16.75 +/- 6.21
Iteration 42/500: rewards 20.06 +/- 9.27
Iteration 43/500: rewards 21.31 +/- 10.2
Iteration 44/500: rewards 17.66 +/- 9.43
Iteration 45/500: rewards 20.16 +/- 13.46
Iteration 46/500: rewards 18.47 +/- 7.02
Iteration 47/500: rewards 18.5 +/- 6.61
Iteration 48/500: rewards 17.09 +/- 7.28
Iteration 49/500: rewards 19.53 +/- 8.16
Iteration 50/500: rewards 23.47 +/- 12.59
Iteration 51/500: rewards 19.69 +/- 13.04
Iteration 52/500: rewards 18.81 +/- 7.21
Iteration 53/500: rewards 20.09 +/- 9.0
Iteration 54/500: rewards 19.5 +/- 8.51
Iteration 55/500: rewards 20.94 +/- 12.74
Iteration 56/500: rewards 15.72 +/- 6.22
Iteration 57/500: rewards 17.03 +/- 6.14
Iteration 58/500: rewards 18.44 +/- 8.09
Iteration 59/500: rewards 20.5 +/- 12.45
Iteration 60/500: rewards 19.44 +/- 8.33
Iteration 61/500: rewards 18.69 +/- 10.34
Iteration 62/500: rewards 18.12 +/- 9.55
Iteration 63/500: rewards 19.16 +/- 10.2
Iteration 64/500: rewards 18.88 +/- 9.71
Iteration 65/500: rewards 21.25 +/- 10.64
Iteration 66/500: rewards 19.94 +/- 8.63
Iteration 67/500: rewards 18.53 +/- 8.87
Iteration 68/500: rewards 18.94 +/- 8.44
Iteration 69/500: rewards 21.44 +/- 9.49
Iteration 70/500: rewards 20.69 +/- 11.74
Iteration 71/500: rewards 18.41 +/- 7.46
Iteration 72/500: rewards 16.94 +/- 8.12
Iteration 73/500: rewards 18.09 +/- 9.57
Iteration 74/500: rewards 21.44 +/- 8.46
Iteration 75/500: rewards 18.34 +/- 8.49
Iteration 76/500: rewards 18.06 +/- 7.91
Iteration 77/500: rewards 18.16 +/- 7.54
Iteration 78/500: rewards 18.94 +/- 8.43
Iteration 79/500: rewards 17.66 +/- 5.49
Iteration 80/500: rewards 16.22 +/- 5.89
Iteration 81/500: rewards 18.81 +/- 7.87
Iteration 82/500: rewards 18.38 +/- 7.52
Iteration 83/500: rewards 18.66 +/- 9.43
Iteration 84/500: rewards 15.62 +/- 5.39
Iteration 85/500: rewards 17.22 +/- 7.57
Iteration 86/500: rewards 19.03 +/- 10.33
Iteration 87/500: rewards 19.03 +/- 9.19
Iteration 88/500: rewards 16.84 +/- 6.73
```

```
Iteration 89/500: rewards 16.72 +/- 7.92
Iteration 90/500: rewards 14.94 +/- 6.2
Iteration 91/500: rewards 14.19 +/- 6.29
Iteration 92/500: rewards 16.12 +/- 7.03
Iteration 93/500: rewards 12.84 +/- 4.6
Iteration 94/500: rewards 12.0 +/- 3.08
Iteration 95/500: rewards 11.88 +/- 3.21
Iteration 96/500: rewards 12.88 +/- 5.67
Iteration 97/500: rewards 11.56 +/- 2.93
Iteration 98/500: rewards 10.38 +/- 3.33
Iteration 99/500: rewards 11.06 +/- 2.81
Iteration 100/500: rewards 10.41 +/- 1.78
Iteration 101/500: rewards 10.38 +/- 1.73
Iteration 102/500: rewards 10.25 +/- 2.66
Iteration 103/500: rewards 10.22 +/- 1.43
Iteration 104/500: rewards 9.69 +/- 1.63
Iteration 105/500: rewards 9.72 +/- 1.18
Iteration 106/500: rewards 10.09 +/- 1.51
Iteration 107/500: rewards 9.94 +/- 1.56
Iteration 108/500: rewards 9.47 +/- 1.06
Iteration 109/500: rewards 9.91 +/- 1.68
Iteration 110/500: rewards 9.53 +/- 0.75
Iteration 111/500: rewards 9.88 +/- 1.17
Iteration 112/500: rewards 9.59 +/- 1.11
Iteration 113/500: rewards 9.38 +/- 0.78
Iteration 114/500: rewards 9.59 +/- 0.78
Iteration 115/500: rewards 9.38 +/- 0.78
Iteration 116/500: rewards 9.38 +/- 0.86
Iteration 117/500: rewards 9.53 +/- 0.75
Iteration 118/500: rewards 9.56 +/- 0.83
Iteration 119/500: rewards 9.56 +/- 0.7
Iteration 120/500: rewards 9.31 +/- 1.45
Iteration 121/500: rewards 9.78 +/- 1.19
Iteration 122/500: rewards 9.44 +/- 0.7
Iteration 123/500: rewards 9.81 +/- 1.21
Iteration 124/500: rewards 9.31 +/- 0.68
Iteration 125/500: rewards 9.41 +/- 0.7
Iteration 126/500: rewards 9.38 +/- 0.86
Iteration 127/500: rewards 9.31 +/- 0.85
Iteration 128/500: rewards 9.41 +/- 0.86
Iteration 129/500: rewards 9.38 +/- 0.86
Iteration 130/500: rewards 9.25 +/- 0.87
Iteration 131/500: rewards 9.0 +/- 0.83
Iteration 132/500: rewards 9.25 +/- 0.75
Iteration 133/500: rewards 9.53 +/- 0.71
Iteration 134/500: rewards 9.53 +/- 0.75
Iteration 135/500: rewards 9.59 +/- 0.96
Iteration 136/500: rewards 9.47 +/- 0.61
```

```
Iteration 137/500: rewards 9.62 +/- 0.86
Iteration 138/500: rewards 9.69 +/- 0.95
Iteration 139/500: rewards 9.66 +/- 0.77
Iteration 140/500: rewards 9.41 +/- 0.86
Iteration 141/500: rewards 9.53 +/- 0.66
Iteration 142/500: rewards 9.78 +/- 0.89
Iteration 143/500: rewards 9.22 +/- 0.89
Iteration 144/500: rewards 9.69 +/- 1.33
Iteration 145/500: rewards 9.28 +/- 0.8
Iteration 146/500: rewards 9.41 +/- 0.82
Iteration 147/500: rewards 9.34 +/- 0.85
Iteration 148/500: rewards 9.5 +/- 0.87
Iteration 149/500: rewards 9.62 +/- 0.82
Iteration 150/500: rewards 9.28 +/- 0.76
Iteration 151/500: rewards 9.53 +/- 0.87
Iteration 152/500: rewards 9.12 +/- 0.65
Iteration 153/500: rewards 9.47 +/- 0.9
Iteration 154/500: rewards 9.44 +/- 0.79
Iteration 155/500: rewards 9.47 +/- 0.9
Iteration 156/500: rewards 9.31 +/- 0.88
Iteration 157/500: rewards 9.72 +/- 1.12
Iteration 158/500: rewards 9.28 +/- 0.8
Iteration 159/500: rewards 9.25 +/- 0.83
Iteration 160/500: rewards 9.38 +/- 0.86
Iteration 161/500: rewards 9.25 +/- 0.83
Iteration 162/500: rewards 9.5 +/- 0.71
Iteration 163/500: rewards 9.62 +/- 0.78
Iteration 164/500: rewards 9.59 +/- 0.9
Iteration 165/500: rewards 9.31 +/- 0.77
Iteration 166/500: rewards 9.25 +/- 0.71
Iteration 167/500: rewards 9.5 +/- 0.87
Iteration 168/500: rewards 9.41 +/- 0.74
Iteration 169/500: rewards 9.38 +/- 0.86
Iteration 170/500: rewards 9.47 +/- 0.83
Iteration 171/500: rewards 9.53 +/- 0.97
Iteration 172/500: rewards 9.84 +/- 1.28
Iteration 173/500: rewards 9.34 +/- 0.85
Iteration 174/500: rewards 9.62 +/- 0.74
Iteration 175/500: rewards 9.44 +/- 0.86
Iteration 176/500: rewards 9.19 +/- 1.13
Iteration 177/500: rewards 9.78 +/- 1.11
Iteration 178/500: rewards 9.41 +/- 0.61
Iteration 179/500: rewards 9.28 +/- 0.76
Iteration 180/500: rewards 9.5 +/- 0.79
Iteration 181/500: rewards 9.25 +/- 0.75
Iteration 182/500: rewards 9.75 +/- 1.06
Iteration 183/500: rewards 9.41 +/- 0.86
Iteration 184/500: rewards 9.31 +/- 0.77
```

```
Iteration 185/500: rewards 9.44 +/- 0.86
Iteration 186/500: rewards 9.38 +/- 0.86
Iteration 187/500: rewards 9.53 +/- 0.66
Iteration 188/500: rewards 9.34 +/- 0.81
Iteration 189/500: rewards 9.19 +/- 0.81
Iteration 190/500: rewards 9.56 +/- 0.83
Iteration 191/500: rewards 9.47 +/- 0.9
Iteration 192/500: rewards 9.44 +/- 1.0
Iteration 193/500: rewards 9.12 +/- 0.86
Iteration 194/500: rewards 9.38 +/- 0.7
Iteration 195/500: rewards 9.31 +/- 0.92
Iteration 196/500: rewards 9.34 +/- 0.77
Iteration 197/500: rewards 9.44 +/- 1.06
Iteration 198/500: rewards 9.25 +/- 0.94
Iteration 199/500: rewards 9.56 +/- 1.62
Iteration 200/500: rewards 9.41 +/- 0.86
Iteration 201/500: rewards 9.25 +/- 0.83
Iteration 202/500: rewards 9.38 +/- 0.74
Iteration 203/500: rewards 9.44 +/- 0.79
Iteration 204/500: rewards 9.72 +/- 0.76
Iteration 205/500: rewards 9.59 +/- 0.96
Iteration 206/500: rewards 9.53 +/- 0.75
Iteration 207/500: rewards 9.69 +/- 0.98
Iteration 208/500: rewards 9.38 +/- 0.82
Iteration 209/500: rewards 9.28 +/- 0.67
Iteration 210/500: rewards 9.41 +/- 0.86
Iteration 211/500: rewards 9.47 +/- 0.71
Iteration 212/500: rewards 9.62 +/- 0.7
Iteration 213/500: rewards 9.41 +/- 0.65
Iteration 214/500: rewards 9.5 +/- 0.9
Iteration 215/500: rewards 9.25 +/- 0.83
Iteration 216/500: rewards 9.19 +/- 0.77
Iteration 217/500: rewards 9.22 +/- 0.82
Iteration 218/500: rewards 9.66 +/- 0.69
Iteration 219/500: rewards 9.12 +/- 0.89
Iteration 220/500: rewards 9.34 +/- 0.64
Iteration 221/500: rewards 9.34 +/- 1.05
Iteration 222/500: rewards 9.19 +/- 0.81
Iteration 223/500: rewards 9.34 +/- 0.73
Iteration 224/500: rewards 9.59 +/- 0.9
Iteration 225/500: rewards 9.28 +/- 1.01
Iteration 226/500: rewards 9.28 +/- 0.67
Iteration 227/500: rewards 9.56 +/- 0.93
Iteration 228/500: rewards 9.66 +/- 0.89
Iteration 229/500: rewards 9.47 +/- 0.79
Iteration 230/500: rewards 9.5 +/- 0.83
Iteration 231/500: rewards 9.81 +/- 0.95
Iteration 232/500: rewards 9.22 +/- 0.74
```

```
Iteration 233/500: rewards 9.56 +/- 0.75
Iteration 234/500: rewards 9.44 +/- 0.97
Iteration 235/500: rewards 9.34 +/- 0.64
Iteration 236/500: rewards 9.53 +/- 0.71
Iteration 237/500: rewards 9.56 +/- 0.61
Iteration 238/500: rewards 9.47 +/- 0.93
Iteration 239/500: rewards 9.34 +/- 0.64
Iteration 240/500: rewards 9.41 +/- 0.86
Iteration 241/500: rewards 9.41 +/- 0.86
Iteration 242/500: rewards 9.53 +/- 0.71
Iteration 243/500: rewards 9.72 +/- 0.8
Iteration 244/500: rewards 9.44 +/- 0.83
Iteration 245/500: rewards 9.28 +/- 1.12
Iteration 246/500: rewards 9.22 +/- 0.93
Iteration 247/500: rewards 9.41 +/- 0.7
Iteration 248/500: rewards 9.72 +/- 1.07
Iteration 249/500: rewards 9.5 +/- 0.9
Iteration 250/500: rewards 9.19 +/- 0.68
Iteration 251/500: rewards 9.59 +/- 0.7
Iteration 252/500: rewards 9.28 +/- 0.72
Iteration 253/500: rewards 9.34 +/- 0.81
Iteration 254/500: rewards 9.38 +/- 0.89
Iteration 255/500: rewards 9.41 +/- 0.7
Iteration 256/500: rewards 9.25 +/- 0.71
Iteration 257/500: rewards 9.38 +/- 0.65
Iteration 258/500: rewards 9.53 +/- 0.9
Iteration 259/500: rewards 9.25 +/- 0.87
Iteration 260/500: rewards 9.41 +/- 0.7
Iteration 261/500: rewards 9.34 +/- 0.73
Iteration 262/500: rewards 9.38 +/- 0.65
Iteration 263/500: rewards 9.38 +/- 0.78
Iteration 264/500: rewards 9.16 +/- 0.87
Iteration 265/500: rewards 9.41 +/- 0.7
Iteration 266/500: rewards 9.53 +/- 0.66
Iteration 267/500: rewards 9.25 +/- 0.83
Iteration 268/500: rewards 9.16 +/- 0.71
Iteration 269/500: rewards 9.56 +/- 0.7
Iteration 270/500: rewards 9.5 +/- 0.94
Iteration 271/500: rewards 9.28 +/- 0.8
Iteration 272/500: rewards 9.5 +/- 0.75
Iteration 273/500: rewards 9.5 +/- 0.79
Iteration 274/500: rewards 9.53 +/- 0.79
Iteration 275/500: rewards 9.22 +/- 0.7
Iteration 276/500: rewards 9.22 +/- 0.78
Iteration 277/500: rewards 9.41 +/- 0.74
Iteration 278/500: rewards 9.25 +/- 0.66
Iteration 279/500: rewards 9.47 +/- 0.71
Iteration 280/500: rewards 9.47 +/- 0.97
```

```
Iteration 281/500: rewards 9.31 +/- 0.73
Iteration 282/500: rewards 9.31 +/- 0.68
Iteration 283/500: rewards 9.19 +/- 0.85
Iteration 284/500: rewards 9.28 +/- 0.72
Iteration 285/500: rewards 9.12 +/- 0.7
Iteration 286/500: rewards 9.25 +/- 0.79
Iteration 287/500: rewards 9.34 +/- 0.77
Iteration 288/500: rewards 9.56 +/- 0.93
Iteration 289/500: rewards 9.25 +/- 0.61
Iteration 290/500: rewards 9.31 +/- 0.77
Iteration 291/500: rewards 9.41 +/- 0.65
Iteration 292/500: rewards 9.38 +/- 0.86
Iteration 293/500: rewards 9.25 +/- 0.79
Iteration 294/500: rewards 9.34 +/- 0.59
Iteration 295/500: rewards 9.28 +/- 0.62
Iteration 296/500: rewards 9.5 +/- 0.71
Iteration 297/500: rewards 9.28 +/- 0.8
Iteration 298/500: rewards 9.59 +/- 1.03
Iteration 299/500: rewards 9.22 +/- 0.78
Iteration 300/500: rewards 9.31 +/- 0.77
Iteration 301/500: rewards 9.5 +/- 0.79
Iteration 302/500: rewards 9.38 +/- 0.74
Iteration 303/500: rewards 9.31 +/- 0.81
Iteration 304/500: rewards 9.44 +/- 0.83
Iteration 305/500: rewards 9.56 +/- 1.2
Iteration 306/500: rewards 9.41 +/- 0.7
Iteration 307/500: rewards 9.72 +/- 0.8
Iteration 308/500: rewards 9.34 +/- 0.81
Iteration 309/500: rewards 9.47 +/- 0.71
Iteration 310/500: rewards 9.44 +/- 0.66
Iteration 311/500: rewards 9.41 +/- 0.82
Iteration 312/500: rewards 9.5 +/- 0.83
Iteration 313/500: rewards 9.34 +/- 0.89
Iteration 314/500: rewards 9.38 +/- 0.78
Iteration 315/500: rewards 9.5 +/- 0.94
Iteration 316/500: rewards 9.41 +/- 0.86
Iteration 317/500: rewards 9.44 +/- 0.79
Iteration 318/500: rewards 9.5 +/- 0.79
Iteration 319/500: rewards 9.5 +/- 0.79
Iteration 320/500: rewards 9.38 +/- 0.7
Iteration 321/500: rewards 9.12 +/- 0.86
Iteration 322/500: rewards 9.25 +/- 0.87
Iteration 323/500: rewards 9.38 +/- 0.99
Iteration 324/500: rewards 9.66 +/- 0.81
Iteration 325/500: rewards 9.62 +/- 0.99
Iteration 326/500: rewards 9.38 +/- 0.96
Iteration 327/500: rewards 9.47 +/- 0.66
Iteration 328/500: rewards 9.56 +/- 0.83
```

```
Iteration 329/500: rewards 9.28 +/- 0.84
Iteration 330/500: rewards 9.44 +/- 0.66
Iteration 331/500: rewards 9.34 +/- 0.64
Iteration 332/500: rewards 9.19 +/- 0.73
Iteration 333/500: rewards 9.41 +/- 0.7
Iteration 334/500: rewards 9.31 +/- 0.77
Iteration 335/500: rewards 9.34 +/- 0.81
Iteration 336/500: rewards 9.47 +/- 0.79
Iteration 337/500: rewards 9.16 +/- 0.71
Iteration 338/500: rewards 9.19 +/- 0.68
Iteration 339/500: rewards 9.31 +/- 0.77
Iteration 340/500: rewards 9.31 +/- 0.77
Iteration 341/500: rewards 9.62 +/- 0.7
Iteration 342/500: rewards 9.38 +/- 0.82
Iteration 343/500: rewards 9.28 +/- 0.72
Iteration 344/500: rewards 9.34 +/- 1.02
Iteration 345/500: rewards 9.47 +/- 0.66
Iteration 346/500: rewards 9.44 +/- 0.66
Iteration 347/500: rewards 9.56 +/- 0.79
Iteration 348/500: rewards 9.19 +/- 0.88
Iteration 349/500: rewards 9.28 +/- 0.72
Iteration 350/500: rewards 9.06 +/- 0.79
Iteration 351/500: rewards 9.38 +/- 0.78
Iteration 352/500: rewards 9.44 +/- 0.75
Iteration 353/500: rewards 9.09 +/- 0.8
Iteration 354/500: rewards 9.38 +/- 0.65
Iteration 355/500: rewards 9.62 +/- 0.6
Iteration 356/500: rewards 9.28 +/- 0.67
Iteration 357/500: rewards 9.28 +/- 0.94
Iteration 358/500: rewards 9.38 +/- 0.6
Iteration 359/500: rewards 9.22 +/- 0.7
Iteration 360/500: rewards 9.25 +/- 0.83
Iteration 361/500: rewards 9.12 +/- 0.89
Iteration 362/500: rewards 9.41 +/- 0.74
Iteration 363/500: rewards 9.56 +/- 0.75
Iteration 364/500: rewards 9.44 +/- 0.75
Iteration 365/500: rewards 9.38 +/- 0.96
Iteration 366/500: rewards 8.91 +/- 0.72
Iteration 367/500: rewards 9.28 +/- 0.8
Iteration 368/500: rewards 9.22 +/- 0.74
Iteration 369/500: rewards 9.31 +/- 0.73
Iteration 370/500: rewards 9.16 +/- 0.87
Iteration 371/500: rewards 9.25 +/- 0.79
Iteration 372/500: rewards 9.28 +/- 0.67
Iteration 373/500: rewards 9.53 +/- 1.0
Iteration 374/500: rewards 9.44 +/- 0.66
Iteration 375/500: rewards 9.38 +/- 0.7
Iteration 376/500: rewards 9.25 +/- 0.66
```

```
Iteration 377/500: rewards 9.31 +/- 0.81
Iteration 378/500: rewards 9.31 +/- 0.81
Iteration 379/500: rewards 9.62 +/- 0.65
Iteration 380/500: rewards 9.62 +/- 0.89
Iteration 381/500: rewards 9.44 +/- 0.83
Iteration 382/500: rewards 9.28 +/- 0.72
Iteration 383/500: rewards 9.31 +/- 0.81
Iteration 384/500: rewards 9.5 +/- 0.9
Iteration 385/500: rewards 9.41 +/- 0.7
Iteration 386/500: rewards 9.31 +/- 0.73
Iteration 387/500: rewards 9.41 +/- 0.86
Iteration 388/500: rewards 9.53 +/- 0.66
Iteration 389/500: rewards 9.38 +/- 0.74
Iteration 390/500: rewards 9.34 +/- 0.59
Iteration 391/500: rewards 9.41 +/- 0.78
Iteration 392/500: rewards 9.47 +/- 0.79
Iteration 393/500: rewards 9.56 +/- 0.83
Iteration 394/500: rewards 9.5 +/- 0.66
Iteration 395/500: rewards 9.06 +/- 0.79
Iteration 396/500: rewards 9.34 +/- 0.69
Iteration 397/500: rewards 9.5 +/- 0.66
Iteration 398/500: rewards 9.12 +/- 0.78
Iteration 399/500: rewards 9.56 +/- 0.83
Iteration 400/500: rewards 9.31 +/- 0.68
Iteration 401/500: rewards 9.28 +/- 0.67
Iteration 402/500: rewards 9.31 +/- 0.77
Iteration 403/500: rewards 9.56 +/- 0.66
Iteration 404/500: rewards 9.09 +/- 0.72
Iteration 405/500: rewards 9.5 +/- 0.9
Iteration 406/500: rewards 9.31 +/- 0.77
Iteration 407/500: rewards 9.19 +/- 0.73
Iteration 408/500: rewards 9.44 +/- 0.61
Iteration 409/500: rewards 9.16 +/- 0.79
Iteration 410/500: rewards 9.53 +/- 0.75
Iteration 411/500: rewards 9.56 +/- 0.66
Iteration 412/500: rewards 9.38 +/- 0.7
Iteration 413/500: rewards 9.31 +/- 0.81
Iteration 414/500: rewards 9.62 +/- 0.7
Iteration 415/500: rewards 9.22 +/- 0.74
Iteration 416/500: rewards 9.19 +/- 0.88
Iteration 417/500: rewards 9.47 +/- 0.75
Iteration 418/500: rewards 9.47 +/- 0.71
Iteration 419/500: rewards 9.47 +/- 0.9
Iteration 420/500: rewards 9.34 +/- 0.73
Iteration 421/500: rewards 9.38 +/- 0.65
Iteration 422/500: rewards 9.09 +/- 0.8
Iteration 423/500: rewards 9.62 +/- 0.54
Iteration 424/500: rewards 9.22 +/- 0.89
```

```
Iteration 425/500: rewards 9.5 +/- 0.83
Iteration 426/500: rewards 9.53 +/- 0.75
Iteration 427/500: rewards 9.62 +/- 0.82
Iteration 428/500: rewards 9.28 +/- 0.8
Iteration 429/500: rewards 9.34 +/- 0.69
Iteration 430/500: rewards 9.47 +/- 0.79
Iteration 431/500: rewards 9.16 +/- 0.79
Iteration 432/500: rewards 9.31 +/- 0.58
Iteration 433/500: rewards 9.19 +/- 0.81
Iteration 434/500: rewards 9.44 +/- 0.61
Iteration 435/500: rewards 9.28 +/- 0.62
Iteration 436/500: rewards 9.19 +/- 0.68
Iteration 437/500: rewards 9.38 +/- 0.74
Iteration 438/500: rewards 9.38 +/- 0.86
Iteration 439/500: rewards 9.44 +/- 0.7
Iteration 440/500: rewards 9.53 +/- 0.75
Iteration 441/500: rewards 9.38 +/- 0.82
Iteration 442/500: rewards 9.31 +/- 0.77
Iteration 443/500: rewards 9.28 +/- 0.76
Iteration 444/500: rewards 9.25 +/- 0.79
Iteration 445/500: rewards 9.38 +/- 0.82
Iteration 446/500: rewards 9.34 +/- 0.81
Iteration 447/500: rewards 9.41 +/- 0.65
Iteration 448/500: rewards 9.47 +/- 0.83
Iteration 449/500: rewards 9.56 +/- 0.79
Iteration 450/500: rewards 9.22 +/- 0.78
Iteration 451/500: rewards 9.44 +/- 0.75
Iteration 452/500: rewards 9.34 +/- 0.69
Iteration 453/500: rewards 9.28 +/- 0.8
Iteration 454/500: rewards 9.47 +/- 0.9
Iteration 455/500: rewards 9.31 +/- 0.68
Iteration 456/500: rewards 9.34 +/- 0.89
Iteration 457/500: rewards 9.38 +/- 0.7
Iteration 458/500: rewards 9.31 +/- 0.73
Iteration 459/500: rewards 9.19 +/- 0.81
Iteration 460/500: rewards 9.44 +/- 0.75
Iteration 461/500: rewards 9.41 +/- 0.65
Iteration 462/500: rewards 9.47 +/- 0.66
Iteration 463/500: rewards 9.25 +/- 0.71
Iteration 464/500: rewards 9.31 +/- 0.77
Iteration 465/500: rewards 9.34 +/- 0.81
Iteration 466/500: rewards 9.25 +/- 0.66
Iteration 467/500: rewards 9.5 +/- 0.66
Iteration 468/500: rewards 9.31 +/- 0.68
Iteration 469/500: rewards 9.25 +/- 0.83
Iteration 470/500: rewards 9.38 +/- 0.65
Iteration 471/500: rewards 9.5 +/- 0.66
Iteration 472/500: rewards 9.47 +/- 0.79
```

```
Iteration 473/500: rewards 9.25 +/- 0.66
Iteration 474/500: rewards 9.5 +/- 0.61
Iteration 475/500: rewards 9.5 +/- 0.75
Iteration 476/500: rewards 9.25 +/- 0.71
Iteration 477/500: rewards 9.25 +/- 0.75
Iteration 478/500: rewards 9.47 +/- 0.71
Iteration 479/500: rewards 9.5 +/- 0.75
Iteration 480/500: rewards 9.41 +/- 0.65
Iteration 481/500: rewards 9.53 +/- 0.61
Iteration 482/500: rewards 9.41 +/- 0.74
Iteration 483/500: rewards 9.25 +/- 0.79
Iteration 484/500: rewards 9.41 +/- 0.78
Iteration 485/500: rewards 9.22 +/- 0.7
Iteration 486/500: rewards 9.31 +/- 0.77
Iteration 487/500: rewards 9.47 +/- 0.71
Iteration 488/500: rewards 9.16 +/- 0.79
Iteration 489/500: rewards 9.22 +/- 0.65
Iteration 490/500: rewards 9.25 +/- 0.71
Iteration 491/500: rewards 9.62 +/- 0.74
Iteration 492/500: rewards 9.31 +/- 0.77
Iteration 493/500: rewards 9.34 +/- 0.81
Iteration 494/500: rewards 9.44 +/- 0.61
Iteration 495/500: rewards 9.38 +/- 0.74
Iteration 496/500: rewards 9.59 +/- 0.7
Iteration 497/500: rewards 9.34 +/- 0.69
Iteration 498/500: rewards 9.41 +/- 0.74
Iteration 499/500: rewards 9.56 +/- 0.75
Iteration 500/500: rewards 9.47 +/- 0.61
The average reward is 9.375
the device is: cpu
--------------------------------------------------
The gamma chosen is: 0.95
The value lr chosen is 0.01
The policy lr chosen is 0.001
Iteration 1/500: rewards 19.0 +/- 7.26
Iteration 2/500: rewards 18.59 +/- 9.64
Iteration 3/500: rewards 17.69 +/- 8.52
Iteration 4/500: rewards 20.94 +/- 6.98
Iteration 5/500: rewards 18.97 +/- 11.74
Iteration 6/500: rewards 18.75 +/- 9.46
Iteration 7/500: rewards 15.97 +/- 5.78
Iteration 8/500: rewards 18.84 +/- 7.36
Iteration 9/500: rewards 20.56 +/- 8.48
Iteration 10/500: rewards 18.84 +/- 6.64
Iteration 11/500: rewards 19.53 +/- 11.07
Iteration 12/500: rewards 20.56 +/- 9.37
Iteration 13/500: rewards 19.62 +/- 10.15
Iteration 14/500: rewards 17.38 +/- 7.72
```

```
Iteration 15/500: rewards 16.72 +/- 6.17
Iteration 16/500: rewards 19.47 +/- 9.12
Iteration 17/500: rewards 19.56 +/- 9.58
Iteration 18/500: rewards 19.25 +/- 9.8
Iteration 19/500: rewards 19.88 +/- 11.06
Iteration 20/500: rewards 17.53 +/- 6.32
Iteration 21/500: rewards 17.16 +/- 7.89
Iteration 22/500: rewards 17.69 +/- 8.68
Iteration 23/500: rewards 18.53 +/- 9.93
Iteration 24/500: rewards 18.94 +/- 8.29
Iteration 25/500: rewards 16.75 +/- 6.73
Iteration 26/500: rewards 17.97 +/- 7.99
Iteration 27/500: rewards 19.38 +/- 9.32
Iteration 28/500: rewards 17.25 +/- 6.94
Iteration 29/500: rewards 18.62 +/- 8.34
Iteration 30/500: rewards 19.09 +/- 12.48
Iteration 31/500: rewards 17.59 +/- 4.94
Iteration 32/500: rewards 17.91 +/- 8.15
Iteration 33/500: rewards 16.91 +/- 6.51
Iteration 34/500: rewards 18.06 +/- 9.23
Iteration 35/500: rewards 17.12 +/- 6.02
Iteration 36/500: rewards 18.22 +/- 8.74
Iteration 37/500: rewards 19.12 +/- 15.02
Iteration 38/500: rewards 17.81 +/- 6.93
Iteration 39/500: rewards 18.88 +/- 12.3
Iteration 40/500: rewards 19.78 +/- 8.98
Iteration 41/500: rewards 15.81 +/- 7.17
Iteration 42/500: rewards 18.44 +/- 7.24
Iteration 43/500: rewards 20.19 +/- 10.17
Iteration 44/500: rewards 18.16 +/- 11.4
Iteration 45/500: rewards 18.56 +/- 7.3
Iteration 46/500: rewards 17.94 +/- 7.1
Iteration 47/500: rewards 20.81 +/- 13.74
Iteration 48/500: rewards 19.09 +/- 7.8
Iteration 49/500: rewards 21.09 +/- 10.42
Iteration 50/500: rewards 16.06 +/- 5.27
Iteration 51/500: rewards 23.31 +/- 17.56
Iteration 52/500: rewards 21.62 +/- 9.35
Iteration 53/500: rewards 21.94 +/- 14.0
Iteration 54/500: rewards 18.81 +/- 8.29
Iteration 55/500: rewards 21.22 +/- 7.97
Iteration 56/500: rewards 19.44 +/- 8.1
Iteration 57/500: rewards 19.38 +/- 7.61
Iteration 58/500: rewards 21.81 +/- 12.02
Iteration 59/500: rewards 23.84 +/- 10.79
Iteration 60/500: rewards 17.03 +/- 6.99
Iteration 61/500: rewards 19.5 +/- 10.05
Iteration 62/500: rewards 26.31 +/- 19.64
```

```
Iteration 63/500: rewards 23.69 +/- 16.43
Iteration 64/500: rewards 22.28 +/- 12.1
Iteration 65/500: rewards 20.09 +/- 9.44
Iteration 66/500: rewards 20.66 +/- 7.8
Iteration 67/500: rewards 20.91 +/- 9.16
Iteration 68/500: rewards 22.25 +/- 13.15
Iteration 69/500: rewards 22.72 +/- 13.5
Iteration 70/500: rewards 21.5 +/- 9.71
Iteration 71/500: rewards 20.31 +/- 7.56
Iteration 72/500: rewards 24.84 +/- 13.7
Iteration 73/500: rewards 24.84 +/- 13.87
Iteration 74/500: rewards 22.12 +/- 10.78
Iteration 75/500: rewards 21.16 +/- 9.68
Iteration 76/500: rewards 23.03 +/- 10.07
Iteration 77/500: rewards 28.91 +/- 16.15
Iteration 78/500: rewards 25.75 +/- 16.2
Iteration 79/500: rewards 25.72 +/- 15.42
Iteration 80/500: rewards 20.53 +/- 7.0
Iteration 81/500: rewards 23.81 +/- 11.85
Iteration 82/500: rewards 24.09 +/- 10.4
Iteration 83/500: rewards 25.47 +/- 17.37
Iteration 84/500: rewards 23.28 +/- 11.38
Iteration 85/500: rewards 22.28 +/- 11.7
Iteration 86/500: rewards 28.25 +/- 17.17
Iteration 87/500: rewards 25.31 +/- 12.43
Iteration 88/500: rewards 26.59 +/- 15.96
Iteration 89/500: rewards 26.47 +/- 17.77
Iteration 90/500: rewards 25.03 +/- 10.2
Iteration 91/500: rewards 24.62 +/- 10.17
Iteration 92/500: rewards 27.28 +/- 14.69
Iteration 93/500: rewards 25.91 +/- 11.4
Iteration 94/500: rewards 23.03 +/- 10.81
Iteration 95/500: rewards 29.59 +/- 17.69
Iteration 96/500: rewards 28.16 +/- 14.78
Iteration 97/500: rewards 27.03 +/- 15.03
Iteration 98/500: rewards 27.5 +/- 13.76
Iteration 99/500: rewards 24.53 +/- 12.93
Iteration 100/500: rewards 28.94 +/- 13.11
Iteration 101/500: rewards 29.03 +/- 14.36
Iteration 102/500: rewards 32.78 +/- 22.08
Iteration 103/500: rewards 31.56 +/- 15.82
Iteration 104/500: rewards 30.41 +/- 19.17
Iteration 105/500: rewards 32.97 +/- 19.7
Iteration 106/500: rewards 23.31 +/- 13.08
Iteration 107/500: rewards 28.5 +/- 17.87
Iteration 108/500: rewards 32.38 +/- 19.39
Iteration 109/500: rewards 31.81 +/- 16.11
Iteration 110/500: rewards 30.59 +/- 20.65
```

```
Iteration 111/500: rewards 34.84 +/- 18.06
Iteration 112/500: rewards 27.38 +/- 15.01
Iteration 113/500: rewards 33.91 +/- 19.96
Iteration 114/500: rewards 33.19 +/- 20.9
Iteration 115/500: rewards 36.56 +/- 22.82
Iteration 116/500: rewards 36.06 +/- 17.24
Iteration 117/500: rewards 28.5 +/- 12.58
Iteration 118/500: rewards 30.56 +/- 12.55
Iteration 119/500: rewards 32.47 +/- 20.12
Iteration 120/500: rewards 45.88 +/- 17.28
Iteration 121/500: rewards 39.16 +/- 17.12
Iteration 122/500: rewards 44.97 +/- 22.61
Iteration 123/500: rewards 36.16 +/- 19.65
Iteration 124/500: rewards 37.06 +/- 25.06
Iteration 125/500: rewards 37.47 +/- 17.98
Iteration 126/500: rewards 37.34 +/- 24.3
Iteration 127/500: rewards 40.62 +/- 24.42
Iteration 128/500: rewards 38.5 +/- 18.76
Iteration 129/500: rewards 38.44 +/- 19.23
Iteration 130/500: rewards 40.28 +/- 18.94
Iteration 131/500: rewards 42.03 +/- 22.68
Iteration 132/500: rewards 41.97 +/- 24.85
Iteration 133/500: rewards 43.53 +/- 28.84
Iteration 134/500: rewards 37.25 +/- 22.06
Iteration 135/500: rewards 40.91 +/- 25.41
Iteration 136/500: rewards 41.59 +/- 18.89
Iteration 137/500: rewards 48.84 +/- 31.39
Iteration 138/500: rewards 47.38 +/- 24.42
Iteration 139/500: rewards 46.97 +/- 28.05
Iteration 140/500: rewards 44.19 +/- 23.63
Iteration 141/500: rewards 42.66 +/- 14.4
Iteration 142/500: rewards 55.34 +/- 28.12
Iteration 143/500: rewards 45.47 +/- 26.44
Iteration 144/500: rewards 47.53 +/- 31.08
Iteration 145/500: rewards 47.38 +/- 26.98
Iteration 146/500: rewards 50.91 +/- 25.79
Iteration 147/500: rewards 61.44 +/- 30.21
Iteration 148/500: rewards 60.75 +/- 29.33
Iteration 149/500: rewards 57.38 +/- 33.06
Iteration 150/500: rewards 46.03 +/- 21.7
Iteration 151/500: rewards 53.06 +/- 27.57
Iteration 152/500: rewards 58.72 +/- 32.65
Iteration 153/500: rewards 52.75 +/- 28.24
Iteration 154/500: rewards 58.84 +/- 23.96
Iteration 155/500: rewards 45.88 +/- 21.24
Iteration 156/500: rewards 41.72 +/- 19.47
Iteration 157/500: rewards 57.16 +/- 33.95
Iteration 158/500: rewards 54.38 +/- 17.47
```

```
Iteration 159/500: rewards 62.94 +/- 27.06
Iteration 160/500: rewards 53.0 +/- 21.32
Iteration 161/500: rewards 57.19 +/- 29.7
Iteration 162/500: rewards 63.41 +/- 34.48
Iteration 163/500: rewards 58.94 +/- 29.16
Iteration 164/500: rewards 63.53 +/- 23.58
Iteration 165/500: rewards 66.66 +/- 38.41
Iteration 166/500: rewards 54.94 +/- 19.92
Iteration 167/500: rewards 61.97 +/- 32.54
Iteration 168/500: rewards 61.66 +/- 17.87
Iteration 169/500: rewards 69.81 +/- 34.37
Iteration 170/500: rewards 58.34 +/- 21.85
Iteration 171/500: rewards 65.34 +/- 32.76
Iteration 172/500: rewards 72.94 +/- 29.65
Iteration 173/500: rewards 69.97 +/- 31.25
Iteration 174/500: rewards 57.47 +/- 23.75
Iteration 175/500: rewards 65.72 +/- 33.34
Iteration 176/500: rewards 76.78 +/- 37.61
Iteration 177/500: rewards 57.06 +/- 24.4
Iteration 178/500: rewards 66.75 +/- 23.51
Iteration 179/500: rewards 63.44 +/- 25.47
Iteration 180/500: rewards 69.44 +/- 32.43
Iteration 181/500: rewards 78.69 +/- 34.19
Iteration 182/500: rewards 77.22 +/- 29.15
Iteration 183/500: rewards 73.72 +/- 31.73
Iteration 184/500: rewards 62.28 +/- 21.3
Iteration 185/500: rewards 81.72 +/- 33.77
Iteration 186/500: rewards 73.0 +/- 25.33
Iteration 187/500: rewards 76.94 +/- 38.63
Iteration 188/500: rewards 72.78 +/- 20.69
Iteration 189/500: rewards 89.56 +/- 33.49
Iteration 190/500: rewards 73.75 +/- 31.64
Iteration 191/500: rewards 67.12 +/- 29.08
Iteration 192/500: rewards 73.0 +/- 24.46
Iteration 193/500: rewards 77.69 +/- 29.43
Iteration 194/500: rewards 87.59 +/- 32.0
Iteration 195/500: rewards 84.72 +/- 31.59
Iteration 196/500: rewards 85.5 +/- 38.08
Iteration 197/500: rewards 79.16 +/- 25.56
Iteration 198/500: rewards 81.06 +/- 36.21
Iteration 199/500: rewards 77.31 +/- 31.44
Iteration 200/500: rewards 83.75 +/- 45.82
Iteration 201/500: rewards 86.88 +/- 36.53
Iteration 202/500: rewards 77.19 +/- 25.49
Iteration 203/500: rewards 90.91 +/- 38.44
Iteration 204/500: rewards 83.06 +/- 33.58
Iteration 205/500: rewards 89.75 +/- 30.37
Iteration 206/500: rewards 95.38 +/- 36.23
```

```
Iteration 207/500: rewards 100.44 +/- 41.59
Iteration 208/500: rewards 87.56 +/- 31.23
Iteration 209/500: rewards 88.91 +/- 30.19
Iteration 210/500: rewards 93.16 +/- 31.65
Iteration 211/500: rewards 99.38 +/- 38.68
Iteration 212/500: rewards 94.47 +/- 33.03
Iteration 213/500: rewards 94.78 +/- 32.48
Iteration 214/500: rewards 99.56 +/- 40.88
Iteration 215/500: rewards 89.38 +/- 37.17
Iteration 216/500: rewards 102.66 +/- 46.44
Iteration 217/500: rewards 105.94 +/- 44.25
Iteration 218/500: rewards 101.0 +/- 27.38
Iteration 219/500: rewards 109.25 +/- 41.52
Iteration 220/500: rewards 115.53 +/- 40.38
Iteration 221/500: rewards 121.94 +/- 49.39
Iteration 222/500: rewards 118.84 +/- 39.39
Iteration 223/500: rewards 108.88 +/- 28.21
Iteration 224/500: rewards 122.84 +/- 45.45
Iteration 225/500: rewards 136.25 +/- 55.3
Iteration 226/500: rewards 152.78 +/- 60.19
Iteration 227/500: rewards 141.75 +/- 45.17
Iteration 228/500: rewards 134.62 +/- 63.06
Iteration 229/500: rewards 162.25 +/- 59.37
Iteration 230/500: rewards 180.09 +/- 91.44
Iteration 231/500: rewards 160.91 +/- 52.97
Iteration 232/500: rewards 161.28 +/- 51.04
Iteration 233/500: rewards 168.56 +/- 62.42
Iteration 234/500: rewards 176.91 +/- 63.69
Iteration 235/500: rewards 196.72 +/- 99.1
Iteration 236/500: rewards 200.91 +/- 84.9
Iteration 237/500: rewards 207.72 +/- 94.41
Iteration 238/500: rewards 225.47 +/- 102.39
Iteration 239/500: rewards 264.72 +/- 103.33
Iteration 240/500: rewards 237.97 +/- 114.77
Iteration 241/500: rewards 206.41 +/- 88.26
Iteration 242/500: rewards 265.62 +/- 108.15
Iteration 243/500: rewards 285.0 +/- 129.47
Iteration 244/500: rewards 335.56 +/- 104.73
Iteration 245/500: rewards 287.62 +/- 116.35
Iteration 246/500: rewards 340.81 +/- 127.32
Iteration 247/500: rewards 366.16 +/- 120.97
Iteration 248/500: rewards 365.53 +/- 136.86
Iteration 249/500: rewards 406.81 +/- 121.13
Iteration 250/500: rewards 378.16 +/- 121.66
Iteration 251/500: rewards 361.94 +/- 120.6
Iteration 252/500: rewards 328.56 +/- 113.56
Iteration 253/500: rewards 304.19 +/- 99.82
Iteration 254/500: rewards 304.19 +/- 115.8
```

```
Iteration 255/500: rewards 319.31 +/- 102.23
Iteration 256/500: rewards 286.19 +/- 109.36
Iteration 257/500: rewards 289.34 +/- 93.61
Iteration 258/500: rewards 297.25 +/- 87.64
Iteration 259/500: rewards 242.66 +/- 89.83
Iteration 260/500: rewards 255.25 +/- 80.55
Iteration 261/500: rewards 233.19 +/- 55.79
Iteration 262/500: rewards 224.5 +/- 69.01
Iteration 263/500: rewards 209.22 +/- 57.14
Iteration 264/500: rewards 194.06 +/- 40.64
Iteration 265/500: rewards 185.56 +/- 38.76
Iteration 266/500: rewards 175.94 +/- 30.06
Iteration 267/500: rewards 195.19 +/- 64.92
Iteration 268/500: rewards 176.34 +/- 44.6
Iteration 269/500: rewards 168.34 +/- 29.59
Iteration 270/500: rewards 182.72 +/- 45.99
Iteration 271/500: rewards 180.12 +/- 36.44
Iteration 272/500: rewards 193.38 +/- 60.42
Iteration 273/500: rewards 186.44 +/- 51.9
Iteration 274/500: rewards 181.06 +/- 30.83
Iteration 275/500: rewards 187.34 +/- 45.19
Iteration 276/500: rewards 185.16 +/- 47.75
Iteration 277/500: rewards 188.03 +/- 50.01
Iteration 278/500: rewards 174.66 +/- 38.58
Iteration 279/500: rewards 183.75 +/- 35.83
Iteration 280/500: rewards 179.75 +/- 36.0
Iteration 281/500: rewards 177.28 +/- 28.4
Iteration 282/500: rewards 180.34 +/- 33.32
Iteration 283/500: rewards 190.06 +/- 49.51
Iteration 284/500: rewards 191.22 +/- 44.8
Iteration 285/500: rewards 195.88 +/- 51.6
Iteration 286/500: rewards 204.84 +/- 59.23
Iteration 287/500: rewards 216.53 +/- 57.84
Iteration 288/500: rewards 221.97 +/- 58.69
Iteration 289/500: rewards 246.78 +/- 81.39
Iteration 290/500: rewards 227.12 +/- 59.08
Iteration 291/500: rewards 248.94 +/- 79.41
Iteration 292/500: rewards 237.09 +/- 78.28
Iteration 293/500: rewards 232.16 +/- 56.77
Iteration 294/500: rewards 237.44 +/- 59.26
Iteration 295/500: rewards 243.06 +/- 71.49
Iteration 296/500: rewards 275.78 +/- 100.22
Iteration 297/500: rewards 263.06 +/- 87.53
Iteration 298/500: rewards 249.75 +/- 68.9
Iteration 299/500: rewards 225.38 +/- 78.12
Iteration 300/500: rewards 235.75 +/- 82.92
Iteration 301/500: rewards 257.34 +/- 96.44
Iteration 302/500: rewards 234.41 +/- 76.97
```

```
Iteration 303/500: rewards 221.62 +/- 59.28
Iteration 304/500: rewards 225.59 +/- 64.38
Iteration 305/500: rewards 268.69 +/- 97.19
Iteration 306/500: rewards 245.81 +/- 84.74
Iteration 307/500: rewards 278.03 +/- 107.11
Iteration 308/500: rewards 240.88 +/- 100.16
Iteration 309/500: rewards 235.94 +/- 56.75
Iteration 310/500: rewards 278.16 +/- 100.81
Iteration 311/500: rewards 260.44 +/- 91.45
Iteration 312/500: rewards 262.06 +/- 66.79
Iteration 313/500: rewards 253.47 +/- 93.95
Iteration 314/500: rewards 237.47 +/- 75.6
Iteration 315/500: rewards 252.0 +/- 88.94
Iteration 316/500: rewards 264.28 +/- 85.88
Iteration 317/500: rewards 250.91 +/- 76.5
Iteration 318/500: rewards 260.5 +/- 96.97
Iteration 319/500: rewards 235.75 +/- 66.5
Iteration 320/500: rewards 242.0 +/- 71.29
Iteration 321/500: rewards 227.88 +/- 65.61
Iteration 322/500: rewards 243.31 +/- 80.77
Iteration 323/500: rewards 237.97 +/- 59.26
Iteration 324/500: rewards 261.12 +/- 95.72
Iteration 325/500: rewards 249.59 +/- 98.03
Iteration 326/500: rewards 236.41 +/- 79.28
Iteration 327/500: rewards 238.22 +/- 79.49
Iteration 328/500: rewards 250.41 +/- 90.43
Iteration 329/500: rewards 236.91 +/- 95.59
Iteration 330/500: rewards 257.66 +/- 93.31
Iteration 331/500: rewards 248.28 +/- 86.58
Iteration 332/500: rewards 241.12 +/- 76.34
Iteration 333/500: rewards 242.84 +/- 60.29
Iteration 334/500: rewards 248.66 +/- 78.71
Iteration 335/500: rewards 257.94 +/- 96.49
Iteration 336/500: rewards 249.47 +/- 84.96
Iteration 337/500: rewards 228.94 +/- 66.47
Iteration 338/500: rewards 240.44 +/- 87.34
Iteration 339/500: rewards 236.03 +/- 80.3
Iteration 340/500: rewards 234.31 +/- 76.26
Iteration 341/500: rewards 229.75 +/- 94.37
Iteration 342/500: rewards 234.34 +/- 79.9
Iteration 343/500: rewards 237.69 +/- 75.51
Iteration 344/500: rewards 223.19 +/- 69.91
Iteration 345/500: rewards 238.75 +/- 91.2
Iteration 346/500: rewards 218.94 +/- 83.63
Iteration 347/500: rewards 200.66 +/- 68.26
Iteration 348/500: rewards 215.94 +/- 62.21
Iteration 349/500: rewards 212.5 +/- 88.11
Iteration 350/500: rewards 238.41 +/- 96.25
```

```
Iteration 351/500: rewards 221.31 +/- 102.55
Iteration 352/500: rewards 194.94 +/- 68.39
Iteration 353/500: rewards 229.34 +/- 83.97
Iteration 354/500: rewards 216.62 +/- 83.86
Iteration 355/500: rewards 189.78 +/- 64.72
Iteration 356/500: rewards 197.28 +/- 63.24
Iteration 357/500: rewards 207.38 +/- 88.33
Iteration 358/500: rewards 214.47 +/- 84.77
Iteration 359/500: rewards 209.41 +/- 64.99
Iteration 360/500: rewards 200.78 +/- 53.8
Iteration 361/500: rewards 237.53 +/- 98.4
Iteration 362/500: rewards 197.28 +/- 72.85
Iteration 363/500: rewards 216.91 +/- 77.89
Iteration 364/500: rewards 198.88 +/- 58.74
Iteration 365/500: rewards 208.12 +/- 66.98
Iteration 366/500: rewards 221.59 +/- 53.93
Iteration 367/500: rewards 189.22 +/- 51.43
Iteration 368/500: rewards 226.66 +/- 74.35
Iteration 369/500: rewards 204.72 +/- 55.58
Iteration 370/500: rewards 216.59 +/- 74.47
Iteration 371/500: rewards 212.0 +/- 62.14
Iteration 372/500: rewards 211.31 +/- 84.01
Iteration 373/500: rewards 190.12 +/- 70.49
Iteration 374/500: rewards 204.09 +/- 67.54
Iteration 375/500: rewards 197.84 +/- 66.82
Iteration 376/500: rewards 204.56 +/- 67.46
Iteration 377/500: rewards 191.31 +/- 56.46
Iteration 378/500: rewards 197.59 +/- 63.97
Iteration 379/500: rewards 200.16 +/- 80.07
Iteration 380/500: rewards 203.19 +/- 70.54
Iteration 381/500: rewards 222.38 +/- 86.26
Iteration 382/500: rewards 211.66 +/- 69.68
Iteration 383/500: rewards 195.94 +/- 67.17
Iteration 384/500: rewards 192.84 +/- 61.36
Iteration 385/500: rewards 199.34 +/- 72.95
Iteration 386/500: rewards 247.03 +/- 88.68
Iteration 387/500: rewards 203.09 +/- 67.93
Iteration 388/500: rewards 218.94 +/- 81.44
Iteration 389/500: rewards 251.59 +/- 93.87
Iteration 390/500: rewards 228.59 +/- 80.29
Iteration 391/500: rewards 192.47 +/- 68.54
Iteration 392/500: rewards 227.84 +/- 62.24
Iteration 393/500: rewards 241.34 +/- 98.35
Iteration 394/500: rewards 239.97 +/- 78.08
Iteration 395/500: rewards 231.62 +/- 82.61
Iteration 396/500: rewards 244.09 +/- 73.33
Iteration 397/500: rewards 226.94 +/- 76.11
Iteration 398/500: rewards 229.03 +/- 77.18
```

```
Iteration 399/500: rewards 206.5 +/- 74.75
Iteration 400/500: rewards 217.0 +/- 60.39
Iteration 401/500: rewards 244.34 +/- 108.95
Iteration 402/500: rewards 218.44 +/- 88.15
Iteration 403/500: rewards 220.97 +/- 85.51
Iteration 404/500: rewards 234.97 +/- 74.64
Iteration 405/500: rewards 200.72 +/- 50.97
Iteration 406/500: rewards 219.81 +/- 79.69
Iteration 407/500: rewards 224.66 +/- 81.49
Iteration 408/500: rewards 248.44 +/- 93.52
Iteration 409/500: rewards 222.34 +/- 80.29
Iteration 410/500: rewards 222.0 +/- 86.46
Iteration 411/500: rewards 235.28 +/- 94.56
Iteration 412/500: rewards 245.06 +/- 112.29
Iteration 413/500: rewards 268.47 +/- 104.62
Iteration 414/500: rewards 232.22 +/- 76.85
Iteration 415/500: rewards 252.88 +/- 89.92
Iteration 416/500: rewards 236.72 +/- 99.85
Iteration 417/500: rewards 242.47 +/- 103.47
Iteration 418/500: rewards 229.75 +/- 70.59
Iteration 419/500: rewards 224.12 +/- 80.4
Iteration 420/500: rewards 252.44 +/- 81.6
Iteration 421/500: rewards 240.47 +/- 67.03
Iteration 422/500: rewards 261.44 +/- 81.76
Iteration 423/500: rewards 237.19 +/- 92.51
Iteration 424/500: rewards 246.38 +/- 74.33
Iteration 425/500: rewards 233.09 +/- 89.8
Iteration 426/500: rewards 264.59 +/- 105.5
Iteration 427/500: rewards 230.53 +/- 81.57
Iteration 428/500: rewards 286.69 +/- 117.17
Iteration 429/500: rewards 225.84 +/- 84.62
Iteration 430/500: rewards 256.91 +/- 98.37
Iteration 431/500: rewards 281.66 +/- 111.88
Iteration 432/500: rewards 284.03 +/- 107.25
Iteration 433/500: rewards 269.47 +/- 92.64
Iteration 434/500: rewards 260.34 +/- 89.4
Iteration 435/500: rewards 305.78 +/- 109.16
Iteration 436/500: rewards 276.41 +/- 91.03
Iteration 437/500: rewards 267.19 +/- 105.28
Iteration 438/500: rewards 290.75 +/- 104.71
Iteration 439/500: rewards 284.59 +/- 119.34
Iteration 440/500: rewards 273.03 +/- 91.86
Iteration 441/500: rewards 285.97 +/- 109.78
Iteration 442/500: rewards 304.44 +/- 126.31
Iteration 443/500: rewards 284.41 +/- 123.66
Iteration 444/500: rewards 312.28 +/- 128.92
Iteration 445/500: rewards 335.31 +/- 119.79
Iteration 446/500: rewards 275.0 +/- 108.73
```

```
Iteration 447/500: rewards 287.47 +/- 112.02
Iteration 448/500: rewards 324.12 +/- 121.89
Iteration 449/500: rewards 312.91 +/- 116.39
Iteration 450/500: rewards 361.12 +/- 137.82
Iteration 451/500: rewards 335.47 +/- 109.38
Iteration 452/500: rewards 336.0 +/- 122.49
Iteration 453/500: rewards 310.91 +/- 123.74
Iteration 454/500: rewards 319.16 +/- 129.23
Iteration 455/500: rewards 358.53 +/- 123.39
Iteration 456/500: rewards 320.78 +/- 124.96
Iteration 457/500: rewards 340.53 +/- 129.56
Iteration 458/500: rewards 347.62 +/- 122.12
Iteration 459/500: rewards 341.88 +/- 113.19
Iteration 460/500: rewards 325.81 +/- 119.67
Iteration 461/500: rewards 346.72 +/- 129.75
Iteration 462/500: rewards 335.16 +/- 115.57
Iteration 463/500: rewards 372.53 +/- 112.13
Iteration 464/500: rewards 370.78 +/- 113.2
Iteration 465/500: rewards 314.72 +/- 110.36
Iteration 466/500: rewards 348.94 +/- 107.14
Iteration 467/500: rewards 365.12 +/- 114.62
Iteration 468/500: rewards 387.75 +/- 117.49
Iteration 469/500: rewards 369.91 +/- 114.61
Iteration 470/500: rewards 331.19 +/- 120.64
Iteration 471/500: rewards 324.44 +/- 114.79
Iteration 472/500: rewards 326.69 +/- 105.85
Iteration 473/500: rewards 330.0 +/- 113.59
Iteration 474/500: rewards 380.22 +/- 106.48
Iteration 475/500: rewards 343.69 +/- 108.43
Iteration 476/500: rewards 346.16 +/- 95.27
Iteration 477/500: rewards 310.88 +/- 111.76
Iteration 478/500: rewards 334.16 +/- 122.43
Iteration 479/500: rewards 317.03 +/- 116.93
Iteration 480/500: rewards 302.31 +/- 98.14
Iteration 481/500: rewards 307.19 +/- 105.43
Iteration 482/500: rewards 313.16 +/- 115.13
Iteration 483/500: rewards 372.47 +/- 114.79
Iteration 484/500: rewards 282.66 +/- 100.84
Iteration 485/500: rewards 313.16 +/- 106.43
Iteration 486/500: rewards 279.12 +/- 93.27
Iteration 487/500: rewards 245.94 +/- 89.75
Iteration 488/500: rewards 332.66 +/- 106.77
Iteration 489/500: rewards 276.84 +/- 83.91
Iteration 490/500: rewards 298.47 +/- 93.06
Iteration 491/500: rewards 234.53 +/- 77.68
Iteration 492/500: rewards 241.69 +/- 85.77
Iteration 493/500: rewards 256.41 +/- 101.73
Iteration 494/500: rewards 231.44 +/- 74.98
```

```
Iteration 495/500: rewards 234.94 +/- 81.99
Iteration 496/500: rewards 225.88 +/- 62.69
Iteration 497/500: rewards 262.91 +/- 98.78
Iteration 498/500: rewards 218.53 +/- 77.29
Iteration 499/500: rewards 208.38 +/- 71.73
Iteration 500/500: rewards 197.66 +/- 54.86
The average reward is 310.58125
the device is: cpu
--------------------------------------------------
The gamma chosen is: 0.95
The value lr chosen is 0.001
The policy lr chosen is 0.001
Iteration 1/500: rewards 19.0 +/- 7.26
Iteration 2/500: rewards 18.59 +/- 9.64
Iteration 3/500: rewards 17.69 +/- 8.52
Iteration 4/500: rewards 20.94 +/- 6.98
Iteration 5/500: rewards 18.97 +/- 11.74
Iteration 6/500: rewards 18.75 +/- 9.46
Iteration 7/500: rewards 17.62 +/- 7.49
Iteration 8/500: rewards 18.59 +/- 7.62
Iteration 9/500: rewards 21.62 +/- 11.03
Iteration 10/500: rewards 21.31 +/- 10.93
Iteration 11/500: rewards 19.22 +/- 8.03
Iteration 12/500: rewards 19.94 +/- 6.98
Iteration 13/500: rewards 16.69 +/- 5.73
Iteration 14/500: rewards 19.72 +/- 9.44
Iteration 15/500: rewards 21.22 +/- 10.57
Iteration 16/500: rewards 17.94 +/- 7.39
Iteration 17/500: rewards 18.47 +/- 8.76
Iteration 18/500: rewards 16.91 +/- 8.02
Iteration 19/500: rewards 18.38 +/- 7.51
Iteration 20/500: rewards 21.0 +/- 10.07
Iteration 21/500: rewards 20.81 +/- 10.41
Iteration 22/500: rewards 23.09 +/- 17.19
Iteration 23/500: rewards 19.31 +/- 9.64
Iteration 24/500: rewards 17.66 +/- 10.07
Iteration 25/500: rewards 17.12 +/- 8.33
Iteration 26/500: rewards 16.78 +/- 7.83
Iteration 27/500: rewards 18.91 +/- 8.64
Iteration 28/500: rewards 23.78 +/- 18.55
Iteration 29/500: rewards 20.47 +/- 7.66
Iteration 30/500: rewards 19.72 +/- 9.08
Iteration 31/500: rewards 17.81 +/- 9.06
Iteration 32/500: rewards 19.56 +/- 8.92
Iteration 33/500: rewards 19.31 +/- 8.52
Iteration 34/500: rewards 21.5 +/- 11.63
Iteration 35/500: rewards 18.62 +/- 6.58
Iteration 36/500: rewards 18.81 +/- 8.38
```

```
Iteration 37/500: rewards 18.72 +/- 8.57
Iteration 38/500: rewards 19.5 +/- 10.41
Iteration 39/500: rewards 15.38 +/- 4.28
Iteration 40/500: rewards 19.16 +/- 8.25
Iteration 41/500: rewards 17.34 +/- 6.83
Iteration 42/500: rewards 20.28 +/- 9.1
Iteration 43/500: rewards 16.78 +/- 6.06
Iteration 44/500: rewards 19.75 +/- 10.9
Iteration 45/500: rewards 21.62 +/- 9.68
Iteration 46/500: rewards 20.44 +/- 10.34
Iteration 47/500: rewards 24.5 +/- 15.4
Iteration 48/500: rewards 20.81 +/- 11.59
Iteration 49/500: rewards 20.38 +/- 12.35
Iteration 50/500: rewards 18.84 +/- 6.96
Iteration 51/500: rewards 20.84 +/- 11.47
Iteration 52/500: rewards 19.25 +/- 9.26
Iteration 53/500: rewards 20.75 +/- 8.19
Iteration 54/500: rewards 17.62 +/- 6.64
Iteration 55/500: rewards 17.59 +/- 8.75
Iteration 56/500: rewards 17.16 +/- 6.29
Iteration 57/500: rewards 22.56 +/- 12.44
Iteration 58/500: rewards 18.69 +/- 6.89
Iteration 59/500: rewards 17.31 +/- 7.7
Iteration 60/500: rewards 16.94 +/- 6.65
Iteration 61/500: rewards 18.5 +/- 7.42
Iteration 62/500: rewards 21.53 +/- 10.54
Iteration 63/500: rewards 18.5 +/- 5.51
Iteration 64/500: rewards 21.25 +/- 13.48
Iteration 65/500: rewards 16.41 +/- 6.72
Iteration 66/500: rewards 19.06 +/- 10.3
Iteration 67/500: rewards 19.31 +/- 8.49
Iteration 68/500: rewards 18.38 +/- 8.23
Iteration 69/500: rewards 16.44 +/- 5.96
Iteration 70/500: rewards 20.34 +/- 10.23
Iteration 71/500: rewards 14.94 +/- 4.84
Iteration 72/500: rewards 22.28 +/- 11.5
Iteration 73/500: rewards 16.03 +/- 7.25
Iteration 74/500: rewards 20.41 +/- 10.79
Iteration 75/500: rewards 19.25 +/- 8.96
Iteration 76/500: rewards 21.12 +/- 10.4
Iteration 77/500: rewards 17.97 +/- 8.87
Iteration 78/500: rewards 18.53 +/- 10.34
Iteration 79/500: rewards 17.34 +/- 7.49
Iteration 80/500: rewards 20.62 +/- 12.28
Iteration 81/500: rewards 18.44 +/- 9.45
Iteration 82/500: rewards 18.31 +/- 11.32
Iteration 83/500: rewards 17.84 +/- 8.24
Iteration 84/500: rewards 20.81 +/- 10.82
```

```
Iteration 85/500: rewards 21.0 +/- 11.97
Iteration 86/500: rewards 16.25 +/- 8.1
Iteration 87/500: rewards 19.66 +/- 10.85
Iteration 88/500: rewards 17.91 +/- 7.6
Iteration 89/500: rewards 17.66 +/- 9.15
Iteration 90/500: rewards 21.06 +/- 9.33
Iteration 91/500: rewards 21.59 +/- 9.27
Iteration 92/500: rewards 15.94 +/- 8.76
Iteration 93/500: rewards 19.25 +/- 8.37
Iteration 94/500: rewards 15.91 +/- 10.52
Iteration 95/500: rewards 15.0 +/- 4.37
Iteration 96/500: rewards 18.41 +/- 8.13
Iteration 97/500: rewards 17.56 +/- 8.33
Iteration 98/500: rewards 17.25 +/- 5.32
Iteration 99/500: rewards 16.22 +/- 6.2
Iteration 100/500: rewards 18.75 +/- 8.8
Iteration 101/500: rewards 18.59 +/- 9.76
Iteration 102/500: rewards 20.62 +/- 13.53
Iteration 103/500: rewards 17.94 +/- 6.75
Iteration 104/500: rewards 18.44 +/- 8.55
Iteration 105/500: rewards 18.19 +/- 9.3
Iteration 106/500: rewards 18.81 +/- 12.01
Iteration 107/500: rewards 16.53 +/- 6.36
Iteration 108/500: rewards 18.16 +/- 7.25
Iteration 109/500: rewards 17.25 +/- 6.56
Iteration 110/500: rewards 18.25 +/- 6.44
Iteration 111/500: rewards 16.59 +/- 9.1
Iteration 112/500: rewards 17.19 +/- 8.41
Iteration 113/500: rewards 14.53 +/- 6.35
Iteration 114/500: rewards 18.12 +/- 9.8
Iteration 115/500: rewards 17.75 +/- 8.29
Iteration 116/500: rewards 19.5 +/- 10.4
Iteration 117/500: rewards 15.81 +/- 9.98
Iteration 118/500: rewards 16.66 +/- 8.82
Iteration 119/500: rewards 14.66 +/- 6.15
Iteration 120/500: rewards 17.06 +/- 7.98
Iteration 121/500: rewards 16.22 +/- 6.16
Iteration 122/500: rewards 15.25 +/- 7.75
Iteration 123/500: rewards 16.03 +/- 6.21
Iteration 124/500: rewards 16.31 +/- 6.33
Iteration 125/500: rewards 15.41 +/- 5.53
Iteration 126/500: rewards 15.78 +/- 9.46
Iteration 127/500: rewards 17.38 +/- 7.75
Iteration 128/500: rewards 17.47 +/- 11.26
Iteration 129/500: rewards 17.19 +/- 7.29
Iteration 130/500: rewards 16.91 +/- 7.22
Iteration 131/500: rewards 17.22 +/- 5.45
Iteration 132/500: rewards 17.31 +/- 6.84
```

```
Iteration 133/500: rewards 14.5 +/- 5.41
Iteration 134/500: rewards 16.97 +/- 7.62
Iteration 135/500: rewards 17.53 +/- 9.32
Iteration 136/500: rewards 16.38 +/- 6.74
Iteration 137/500: rewards 16.88 +/- 8.81
Iteration 138/500: rewards 15.84 +/- 7.28
Iteration 139/500: rewards 15.69 +/- 7.45
Iteration 140/500: rewards 16.16 +/- 5.96
Iteration 141/500: rewards 18.28 +/- 9.56
Iteration 142/500: rewards 16.56 +/- 7.52
Iteration 143/500: rewards 17.25 +/- 7.42
Iteration 144/500: rewards 17.72 +/- 11.6
Iteration 145/500: rewards 17.41 +/- 8.38
Iteration 146/500: rewards 17.06 +/- 9.57
Iteration 147/500: rewards 16.34 +/- 7.34
Iteration 148/500: rewards 16.28 +/- 8.3
Iteration 149/500: rewards 17.75 +/- 6.6
Iteration 150/500: rewards 15.44 +/- 5.58
Iteration 151/500: rewards 17.41 +/- 7.29
Iteration 152/500: rewards 15.59 +/- 8.2
Iteration 153/500: rewards 15.53 +/- 4.93
Iteration 154/500: rewards 15.34 +/- 6.73
Iteration 155/500: rewards 14.84 +/- 6.19
Iteration 156/500: rewards 15.59 +/- 5.26
Iteration 157/500: rewards 15.38 +/- 5.88
Iteration 158/500: rewards 13.56 +/- 2.94
Iteration 159/500: rewards 17.97 +/- 8.66
Iteration 160/500: rewards 16.19 +/- 7.67
Iteration 161/500: rewards 16.91 +/- 6.77
Iteration 162/500: rewards 14.72 +/- 5.14
Iteration 163/500: rewards 16.47 +/- 6.52
Iteration 164/500: rewards 14.5 +/- 4.95
Iteration 165/500: rewards 16.47 +/- 6.55
Iteration 166/500: rewards 17.84 +/- 8.7
Iteration 167/500: rewards 17.19 +/- 8.0
Iteration 168/500: rewards 14.88 +/- 3.78
Iteration 169/500: rewards 17.22 +/- 7.16
Iteration 170/500: rewards 15.56 +/- 4.48
Iteration 171/500: rewards 17.91 +/- 8.83
Iteration 172/500: rewards 15.12 +/- 4.73
Iteration 173/500: rewards 15.62 +/- 6.85
Iteration 174/500: rewards 15.53 +/- 6.85
Iteration 175/500: rewards 19.0 +/- 10.06
Iteration 176/500: rewards 14.47 +/- 3.16
Iteration 177/500: rewards 15.78 +/- 8.26
Iteration 178/500: rewards 17.0 +/- 7.71
Iteration 179/500: rewards 16.03 +/- 4.69
Iteration 180/500: rewards 17.06 +/- 9.78
```

```
Iteration 181/500: rewards 16.69 +/- 7.83
Iteration 182/500: rewards 16.78 +/- 8.33
Iteration 183/500: rewards 18.16 +/- 7.77
Iteration 184/500: rewards 18.91 +/- 14.31
Iteration 185/500: rewards 16.78 +/- 6.67
Iteration 186/500: rewards 16.09 +/- 6.67
Iteration 187/500: rewards 15.78 +/- 6.41
Iteration 188/500: rewards 19.81 +/- 10.02
Iteration 189/500: rewards 18.38 +/- 9.89
Iteration 190/500: rewards 16.84 +/- 6.34
Iteration 191/500: rewards 16.75 +/- 6.51
Iteration 192/500: rewards 17.19 +/- 6.31
Iteration 193/500: rewards 18.62 +/- 6.63
Iteration 194/500: rewards 18.75 +/- 9.72
Iteration 195/500: rewards 19.53 +/- 9.98
Iteration 196/500: rewards 18.03 +/- 7.37
Iteration 197/500: rewards 19.06 +/- 8.96
Iteration 198/500: rewards 18.5 +/- 9.62
Iteration 199/500: rewards 17.44 +/- 7.27
Iteration 200/500: rewards 18.94 +/- 9.67
Iteration 201/500: rewards 17.44 +/- 6.68
Iteration 202/500: rewards 18.28 +/- 9.24
Iteration 203/500: rewards 20.03 +/- 7.7
Iteration 204/500: rewards 15.88 +/- 5.98
Iteration 205/500: rewards 20.84 +/- 9.6
Iteration 206/500: rewards 20.72 +/- 8.95
Iteration 207/500: rewards 18.31 +/- 8.84
Iteration 208/500: rewards 22.06 +/- 18.5
Iteration 209/500: rewards 18.38 +/- 9.22
Iteration 210/500: rewards 21.91 +/- 9.35
Iteration 211/500: rewards 21.38 +/- 10.2
Iteration 212/500: rewards 17.09 +/- 6.27
Iteration 213/500: rewards 20.22 +/- 10.37
Iteration 214/500: rewards 21.0 +/- 9.73
Iteration 215/500: rewards 21.28 +/- 10.49
Iteration 216/500: rewards 20.62 +/- 7.99
Iteration 217/500: rewards 21.25 +/- 10.0
Iteration 218/500: rewards 21.5 +/- 9.44
Iteration 219/500: rewards 19.25 +/- 7.31
Iteration 220/500: rewards 26.78 +/- 16.3
Iteration 221/500: rewards 26.69 +/- 19.46
Iteration 222/500: rewards 18.91 +/- 6.79
Iteration 223/500: rewards 23.81 +/- 13.51
Iteration 224/500: rewards 22.97 +/- 13.48
Iteration 225/500: rewards 26.12 +/- 20.34
Iteration 226/500: rewards 27.59 +/- 18.25
Iteration 227/500: rewards 24.25 +/- 14.73
Iteration 228/500: rewards 22.47 +/- 11.93
```

```
Iteration 229/500: rewards 23.62 +/- 11.14
Iteration 230/500: rewards 27.69 +/- 18.97
Iteration 231/500: rewards 20.22 +/- 10.06
Iteration 232/500: rewards 21.44 +/- 8.93
Iteration 233/500: rewards 31.34 +/- 20.93
Iteration 234/500: rewards 30.03 +/- 17.69
Iteration 235/500: rewards 27.62 +/- 18.48
Iteration 236/500: rewards 23.88 +/- 9.64
Iteration 237/500: rewards 30.06 +/- 19.87
Iteration 238/500: rewards 28.25 +/- 18.5
Iteration 239/500: rewards 24.62 +/- 9.11
Iteration 240/500: rewards 30.28 +/- 15.34
Iteration 241/500: rewards 30.78 +/- 15.79
Iteration 242/500: rewards 27.28 +/- 12.78
Iteration 243/500: rewards 30.31 +/- 14.77
Iteration 244/500: rewards 32.97 +/- 18.22
Iteration 245/500: rewards 37.75 +/- 26.81
Iteration 246/500: rewards 32.41 +/- 22.23
Iteration 247/500: rewards 37.41 +/- 19.62
Iteration 248/500: rewards 28.28 +/- 14.47
Iteration 249/500: rewards 36.16 +/- 16.9
Iteration 250/500: rewards 39.41 +/- 22.56
Iteration 251/500: rewards 33.59 +/- 20.2
Iteration 252/500: rewards 40.0 +/- 32.84
Iteration 253/500: rewards 39.53 +/- 22.04
Iteration 254/500: rewards 33.25 +/- 21.13
Iteration 255/500: rewards 34.69 +/- 13.37
Iteration 256/500: rewards 42.28 +/- 21.16
Iteration 257/500: rewards 40.56 +/- 22.4
Iteration 258/500: rewards 37.22 +/- 17.39
Iteration 259/500: rewards 40.56 +/- 32.16
Iteration 260/500: rewards 37.75 +/- 17.99
Iteration 261/500: rewards 38.22 +/- 26.41
Iteration 262/500: rewards 43.78 +/- 26.31
Iteration 263/500: rewards 42.41 +/- 18.85
Iteration 264/500: rewards 47.0 +/- 25.67
Iteration 265/500: rewards 42.19 +/- 21.32
Iteration 266/500: rewards 42.78 +/- 14.48
Iteration 267/500: rewards 56.94 +/- 26.5
Iteration 268/500: rewards 43.97 +/- 21.3
Iteration 269/500: rewards 55.44 +/- 28.3
Iteration 270/500: rewards 50.19 +/- 22.2
Iteration 271/500: rewards 47.25 +/- 23.36
Iteration 272/500: rewards 52.28 +/- 19.68
Iteration 273/500: rewards 49.62 +/- 19.08
Iteration 274/500: rewards 58.56 +/- 36.97
Iteration 275/500: rewards 57.28 +/- 24.51
Iteration 276/500: rewards 54.81 +/- 17.65
```

```
Iteration 277/500: rewards 60.66 +/- 29.89
Iteration 278/500: rewards 53.12 +/- 25.19
Iteration 279/500: rewards 55.78 +/- 17.52
Iteration 280/500: rewards 53.28 +/- 19.48
Iteration 281/500: rewards 59.78 +/- 23.59
Iteration 282/500: rewards 58.69 +/- 20.33
Iteration 283/500: rewards 52.47 +/- 18.96
Iteration 284/500: rewards 62.12 +/- 22.84
Iteration 285/500: rewards 56.66 +/- 17.1
Iteration 286/500: rewards 63.75 +/- 23.35
Iteration 287/500: rewards 61.38 +/- 27.29
Iteration 288/500: rewards 59.5 +/- 19.03
Iteration 289/500: rewards 65.38 +/- 16.49
Iteration 290/500: rewards 62.81 +/- 25.47
Iteration 291/500: rewards 61.66 +/- 25.2
Iteration 292/500: rewards 60.03 +/- 16.86
Iteration 293/500: rewards 62.28 +/- 24.82
Iteration 294/500: rewards 61.56 +/- 22.44
Iteration 295/500: rewards 66.59 +/- 26.95
Iteration 296/500: rewards 59.53 +/- 16.34
Iteration 297/500: rewards 66.88 +/- 23.98
Iteration 298/500: rewards 54.19 +/- 19.59
Iteration 299/500: rewards 60.38 +/- 21.72
Iteration 300/500: rewards 68.34 +/- 25.98
Iteration 301/500: rewards 59.81 +/- 24.23
Iteration 302/500: rewards 61.19 +/- 18.76
Iteration 303/500: rewards 59.44 +/- 17.09
Iteration 304/500: rewards 63.41 +/- 19.69
Iteration 305/500: rewards 67.66 +/- 30.11
Iteration 306/500: rewards 68.59 +/- 28.25
Iteration 307/500: rewards 60.5 +/- 28.41
Iteration 308/500: rewards 66.75 +/- 17.82
Iteration 309/500: rewards 64.72 +/- 17.06
Iteration 310/500: rewards 62.66 +/- 18.38
Iteration 311/500: rewards 67.44 +/- 19.04
Iteration 312/500: rewards 66.25 +/- 17.49
Iteration 313/500: rewards 69.06 +/- 24.43
Iteration 314/500: rewards 68.62 +/- 21.39
Iteration 315/500: rewards 74.0 +/- 25.9
Iteration 316/500: rewards 70.12 +/- 21.81
Iteration 317/500: rewards 71.47 +/- 24.39
Iteration 318/500: rewards 68.91 +/- 19.74
Iteration 319/500: rewards 71.5 +/- 19.41
Iteration 320/500: rewards 74.31 +/- 29.35
Iteration 321/500: rewards 65.88 +/- 21.4
Iteration 322/500: rewards 72.0 +/- 31.63
Iteration 323/500: rewards 70.25 +/- 39.15
Iteration 324/500: rewards 70.72 +/- 21.6
```

```
Iteration 325/500: rewards 70.34 +/- 22.29
Iteration 326/500: rewards 69.34 +/- 23.38
Iteration 327/500: rewards 68.59 +/- 17.13
Iteration 328/500: rewards 75.5 +/- 36.49
Iteration 329/500: rewards 67.44 +/- 21.53
Iteration 330/500: rewards 79.06 +/- 28.31
Iteration 331/500: rewards 80.94 +/- 32.47
Iteration 332/500: rewards 72.09 +/- 25.95
Iteration 333/500: rewards 74.16 +/- 24.76
Iteration 334/500: rewards 76.09 +/- 32.32
Iteration 335/500: rewards 82.44 +/- 34.28
Iteration 336/500: rewards 73.91 +/- 24.36
Iteration 337/500: rewards 79.97 +/- 30.44
Iteration 338/500: rewards 74.31 +/- 25.02
Iteration 339/500: rewards 78.38 +/- 23.89
Iteration 340/500: rewards 82.69 +/- 30.07
Iteration 341/500: rewards 80.25 +/- 28.89
Iteration 342/500: rewards 74.78 +/- 23.95
Iteration 343/500: rewards 78.91 +/- 27.25
Iteration 344/500: rewards 77.41 +/- 30.71
Iteration 345/500: rewards 76.75 +/- 19.09
Iteration 346/500: rewards 90.44 +/- 43.66
Iteration 347/500: rewards 93.28 +/- 37.17
Iteration 348/500: rewards 77.41 +/- 26.75
Iteration 349/500: rewards 80.59 +/- 31.38
Iteration 350/500: rewards 87.47 +/- 37.07
Iteration 351/500: rewards 82.66 +/- 30.25
Iteration 352/500: rewards 86.84 +/- 30.63
Iteration 353/500: rewards 79.16 +/- 24.17
Iteration 354/500: rewards 95.19 +/- 32.78
Iteration 355/500: rewards 90.34 +/- 22.96
Iteration 356/500: rewards 80.0 +/- 23.96
Iteration 357/500: rewards 76.19 +/- 24.79
Iteration 358/500: rewards 87.34 +/- 28.51
Iteration 359/500: rewards 86.81 +/- 33.74
Iteration 360/500: rewards 89.28 +/- 36.21
Iteration 361/500: rewards 87.38 +/- 42.41
Iteration 362/500: rewards 82.19 +/- 27.29
Iteration 363/500: rewards 94.28 +/- 32.38
Iteration 364/500: rewards 89.47 +/- 28.36
Iteration 365/500: rewards 84.78 +/- 25.86
Iteration 366/500: rewards 75.25 +/- 26.64
Iteration 367/500: rewards 79.56 +/- 31.93
Iteration 368/500: rewards 86.75 +/- 44.65
Iteration 369/500: rewards 93.69 +/- 33.73
Iteration 370/500: rewards 81.25 +/- 23.45
Iteration 371/500: rewards 83.53 +/- 23.47
Iteration 372/500: rewards 93.25 +/- 41.89
```

```
Iteration 373/500: rewards 89.59 +/- 28.33
Iteration 374/500: rewards 91.91 +/- 34.01
Iteration 375/500: rewards 90.84 +/- 28.37
Iteration 376/500: rewards 87.0 +/- 28.68
Iteration 377/500: rewards 88.06 +/- 33.4
Iteration 378/500: rewards 81.09 +/- 31.02
Iteration 379/500: rewards 98.25 +/- 25.13
Iteration 380/500: rewards 99.88 +/- 35.37
Iteration 381/500: rewards 97.34 +/- 42.0
Iteration 382/500: rewards 98.03 +/- 42.03
Iteration 383/500: rewards 89.91 +/- 35.27
Iteration 384/500: rewards 91.25 +/- 29.85
Iteration 385/500: rewards 87.09 +/- 35.04
Iteration 386/500: rewards 80.53 +/- 27.38
Iteration 387/500: rewards 88.44 +/- 32.47
Iteration 388/500: rewards 85.88 +/- 23.36
Iteration 389/500: rewards 82.09 +/- 27.9
Iteration 390/500: rewards 85.47 +/- 20.21
Iteration 391/500: rewards 92.91 +/- 36.79
Iteration 392/500: rewards 86.94 +/- 28.28
Iteration 393/500: rewards 94.88 +/- 31.61
Iteration 394/500: rewards 80.81 +/- 24.26
Iteration 395/500: rewards 80.22 +/- 26.41
Iteration 396/500: rewards 79.81 +/- 19.97
Iteration 397/500: rewards 83.06 +/- 25.94
Iteration 398/500: rewards 75.91 +/- 21.59
Iteration 399/500: rewards 84.84 +/- 34.5
Iteration 400/500: rewards 78.44 +/- 25.82
Iteration 401/500: rewards 82.59 +/- 29.54
Iteration 402/500: rewards 80.84 +/- 18.29
Iteration 403/500: rewards 87.5 +/- 27.94
Iteration 404/500: rewards 73.34 +/- 19.2
Iteration 405/500: rewards 83.19 +/- 27.34
Iteration 406/500: rewards 81.44 +/- 25.56
Iteration 407/500: rewards 71.97 +/- 23.94
Iteration 408/500: rewards 76.38 +/- 19.91
Iteration 409/500: rewards 73.34 +/- 17.08
Iteration 410/500: rewards 82.0 +/- 21.52
Iteration 411/500: rewards 81.69 +/- 19.71
Iteration 412/500: rewards 76.19 +/- 18.24
Iteration 413/500: rewards 69.5 +/- 18.38
Iteration 414/500: rewards 71.88 +/- 17.02
Iteration 415/500: rewards 71.25 +/- 22.23
Iteration 416/500: rewards 70.5 +/- 18.11
Iteration 417/500: rewards 71.69 +/- 17.63
Iteration 418/500: rewards 68.12 +/- 15.4
Iteration 419/500: rewards 67.12 +/- 13.39
Iteration 420/500: rewards 67.69 +/- 14.18
```

```
Iteration 421/500: rewards 71.09 +/- 17.33
Iteration 422/500: rewards 64.31 +/- 16.36
Iteration 423/500: rewards 69.38 +/- 17.66
Iteration 424/500: rewards 65.66 +/- 22.6
Iteration 425/500: rewards 68.19 +/- 18.59
Iteration 426/500: rewards 68.44 +/- 21.9
Iteration 427/500: rewards 70.78 +/- 21.26
Iteration 428/500: rewards 65.19 +/- 17.45
Iteration 429/500: rewards 67.44 +/- 19.25
Iteration 430/500: rewards 62.84 +/- 14.22
Iteration 431/500: rewards 61.31 +/- 13.62
Iteration 432/500: rewards 61.66 +/- 13.98
Iteration 433/500: rewards 58.31 +/- 13.15
Iteration 434/500: rewards 64.0 +/- 15.45
Iteration 435/500: rewards 62.41 +/- 11.49
Iteration 436/500: rewards 58.31 +/- 16.14
Iteration 437/500: rewards 59.72 +/- 13.12
Iteration 438/500: rewards 61.09 +/- 14.51
Iteration 439/500: rewards 64.62 +/- 16.18
Iteration 440/500: rewards 62.72 +/- 14.48
Iteration 441/500: rewards 57.62 +/- 13.5
Iteration 442/500: rewards 58.09 +/- 12.05
Iteration 443/500: rewards 61.28 +/- 18.37
Iteration 444/500: rewards 56.91 +/- 15.49
Iteration 445/500: rewards 59.41 +/- 17.99
Iteration 446/500: rewards 56.84 +/- 13.41
Iteration 447/500: rewards 57.72 +/- 15.61
Iteration 448/500: rewards 60.91 +/- 14.49
Iteration 449/500: rewards 60.28 +/- 12.76
Iteration 450/500: rewards 51.81 +/- 12.65
Iteration 451/500: rewards 54.19 +/- 13.35
Iteration 452/500: rewards 56.88 +/- 12.52
Iteration 453/500: rewards 55.03 +/- 16.78
Iteration 454/500: rewards 56.84 +/- 15.27
Iteration 455/500: rewards 53.09 +/- 13.08
Iteration 456/500: rewards 54.56 +/- 13.02
Iteration 457/500: rewards 55.0 +/- 14.8
Iteration 458/500: rewards 52.38 +/- 10.99
Iteration 459/500: rewards 53.16 +/- 13.88
Iteration 460/500: rewards 53.03 +/- 11.38
Iteration 461/500: rewards 54.12 +/- 10.63
Iteration 462/500: rewards 54.28 +/- 10.1
Iteration 463/500: rewards 49.38 +/- 10.65
Iteration 464/500: rewards 50.88 +/- 11.43
Iteration 465/500: rewards 55.12 +/- 14.35
Iteration 466/500: rewards 47.69 +/- 10.83
Iteration 467/500: rewards 52.59 +/- 13.85
Iteration 468/500: rewards 51.12 +/- 12.37
```

```
Iteration 469/500: rewards 51.69 +/- 11.46
Iteration 470/500: rewards 52.41 +/- 13.38
Iteration 471/500: rewards 54.59 +/- 12.21
Iteration 472/500: rewards 54.19 +/- 11.67
Iteration 473/500: rewards 51.75 +/- 13.38
Iteration 474/500: rewards 51.66 +/- 9.93
Iteration 475/500: rewards 51.66 +/- 11.95
Iteration 476/500: rewards 48.78 +/- 11.93
Iteration 477/500: rewards 49.84 +/- 12.29
Iteration 478/500: rewards 50.62 +/- 10.51
Iteration 479/500: rewards 54.03 +/- 13.65
Iteration 480/500: rewards 49.0 +/- 11.7
Iteration 481/500: rewards 51.25 +/- 9.06
Iteration 482/500: rewards 48.72 +/- 11.34
Iteration 483/500: rewards 47.66 +/- 10.81
Iteration 484/500: rewards 49.81 +/- 10.54
Iteration 485/500: rewards 45.0 +/- 8.2
Iteration 486/500: rewards 50.84 +/- 13.68
Iteration 487/500: rewards 51.09 +/- 12.46
Iteration 488/500: rewards 44.66 +/- 9.47
Iteration 489/500: rewards 47.03 +/- 10.86
Iteration 490/500: rewards 45.84 +/- 9.99
Iteration 491/500: rewards 48.72 +/- 9.15
Iteration 492/500: rewards 46.41 +/- 9.23
Iteration 493/500: rewards 48.25 +/- 11.86
Iteration 494/500: rewards 49.09 +/- 10.83
Iteration 495/500: rewards 47.62 +/- 10.77
Iteration 496/500: rewards 49.75 +/- 11.15
Iteration 497/500: rewards 49.34 +/- 9.77
Iteration 498/500: rewards 48.34 +/- 11.17
Iteration 499/500: rewards 49.44 +/- 12.04
Iteration 500/500: rewards 49.75 +/- 11.34
The average reward is 50.96375
the device is: cpu
------------------------------------------------
The gamma chosen is: 0.95
The value lr chosen is 0.0001
The policy lr chosen is 0.001
Iteration 1/500: rewards 19.0 +/- 7.26
Iteration 2/500: rewards 18.59 +/- 9.64
Iteration 3/500: rewards 17.69 +/- 8.52
Iteration 4/500: rewards 20.94 +/- 6.98
Iteration 5/500: rewards 18.97 +/- 11.74
Iteration 6/500: rewards 18.75 +/- 9.46
Iteration 7/500: rewards 17.62 +/- 7.49
Iteration 8/500: rewards 18.59 +/- 7.52
Iteration 9/500: rewards 21.62 +/- 11.03
Iteration 10/500: rewards 21.31 +/- 10.93
```

```
Iteration 11/500: rewards 19.22 +/- 8.03
Iteration 12/500: rewards 19.94 +/- 6.98
Iteration 13/500: rewards 16.69 +/- 5.73
Iteration 14/500: rewards 19.72 +/- 9.44
Iteration 15/500: rewards 21.22 +/- 10.57
Iteration 16/500: rewards 17.94 +/- 7.39
Iteration 17/500: rewards 18.47 +/- 8.76
Iteration 18/500: rewards 16.91 +/- 8.02
Iteration 19/500: rewards 18.12 +/- 8.95
Iteration 20/500: rewards 20.06 +/- 10.26
Iteration 21/500: rewards 19.97 +/- 9.01
Iteration 22/500: rewards 20.09 +/- 10.02
Iteration 23/500: rewards 19.12 +/- 8.1
Iteration 24/500: rewards 17.66 +/- 8.57
Iteration 25/500: rewards 17.53 +/- 8.27
Iteration 26/500: rewards 17.66 +/- 6.78
Iteration 27/500: rewards 20.44 +/- 8.84
Iteration 28/500: rewards 18.06 +/- 7.82
Iteration 29/500: rewards 20.84 +/- 9.96
Iteration 30/500: rewards 19.69 +/- 7.76
Iteration 31/500: rewards 21.31 +/- 13.42
Iteration 32/500: rewards 17.03 +/- 5.4
Iteration 33/500: rewards 20.84 +/- 12.09
Iteration 34/500: rewards 18.31 +/- 9.37
Iteration 35/500: rewards 21.31 +/- 8.16
Iteration 36/500: rewards 17.81 +/- 7.86
Iteration 37/500: rewards 20.53 +/- 11.82
Iteration 38/500: rewards 20.22 +/- 8.73
Iteration 39/500: rewards 16.84 +/- 6.66
Iteration 40/500: rewards 20.94 +/- 12.44
Iteration 41/500: rewards 19.84 +/- 11.99
Iteration 42/500: rewards 23.0 +/- 12.72
Iteration 43/500: rewards 18.72 +/- 8.03
Iteration 44/500: rewards 17.44 +/- 7.3
Iteration 45/500: rewards 18.22 +/- 7.55
Iteration 46/500: rewards 20.88 +/- 13.45
Iteration 47/500: rewards 21.81 +/- 14.84
Iteration 48/500: rewards 21.12 +/- 14.44
Iteration 49/500: rewards 19.84 +/- 10.91
Iteration 50/500: rewards 21.75 +/- 12.65
Iteration 51/500: rewards 22.12 +/- 11.13
Iteration 52/500: rewards 19.88 +/- 12.66
Iteration 53/500: rewards 20.0 +/- 9.07
Iteration 54/500: rewards 19.12 +/- 9.12
Iteration 55/500: rewards 19.75 +/- 13.86
Iteration 56/500: rewards 20.75 +/- 7.58
Iteration 57/500: rewards 18.88 +/- 6.1
Iteration 58/500: rewards 18.97 +/- 8.18
```

```
Iteration 59/500: rewards 17.56 +/- 9.85
Iteration 60/500: rewards 21.0 +/- 10.2
Iteration 61/500: rewards 19.62 +/- 10.48
Iteration 62/500: rewards 19.81 +/- 8.75
Iteration 63/500: rewards 23.53 +/- 13.75
Iteration 64/500: rewards 19.72 +/- 7.41
Iteration 65/500: rewards 20.53 +/- 10.12
Iteration 66/500: rewards 20.75 +/- 10.41
Iteration 67/500: rewards 17.94 +/- 9.77
Iteration 68/500: rewards 16.62 +/- 7.07
Iteration 69/500: rewards 18.56 +/- 8.39
Iteration 70/500: rewards 17.59 +/- 10.23
Iteration 71/500: rewards 20.62 +/- 6.98
Iteration 72/500: rewards 19.78 +/- 11.86
Iteration 73/500: rewards 18.09 +/- 7.0
Iteration 74/500: rewards 20.78 +/- 12.01
Iteration 75/500: rewards 19.94 +/- 8.4
Iteration 76/500: rewards 21.03 +/- 12.0
Iteration 77/500: rewards 19.44 +/- 9.99
Iteration 78/500: rewards 21.16 +/- 10.23
Iteration 79/500: rewards 20.44 +/- 7.11
Iteration 80/500: rewards 20.0 +/- 8.54
Iteration 81/500: rewards 19.78 +/- 8.09
Iteration 82/500: rewards 21.34 +/- 11.68
Iteration 83/500: rewards 18.09 +/- 10.7
Iteration 84/500: rewards 18.75 +/- 9.34
Iteration 85/500: rewards 19.5 +/- 8.9
Iteration 86/500: rewards 19.66 +/- 8.39
Iteration 87/500: rewards 19.78 +/- 9.28
Iteration 88/500: rewards 19.84 +/- 10.86
Iteration 89/500: rewards 19.25 +/- 8.31
Iteration 90/500: rewards 19.22 +/- 10.31
Iteration 91/500: rewards 22.56 +/- 13.07
Iteration 92/500: rewards 19.12 +/- 10.98
Iteration 93/500: rewards 18.72 +/- 7.72
Iteration 94/500: rewards 20.25 +/- 11.61
Iteration 95/500: rewards 18.81 +/- 9.47
Iteration 96/500: rewards 20.72 +/- 10.89
Iteration 97/500: rewards 19.72 +/- 7.33
Iteration 98/500: rewards 22.5 +/- 12.68
Iteration 99/500: rewards 19.94 +/- 9.41
Iteration 100/500: rewards 22.28 +/- 11.89
Iteration 101/500: rewards 19.94 +/- 9.67
Iteration 102/500: rewards 19.09 +/- 8.37
Iteration 103/500: rewards 21.84 +/- 9.84
Iteration 104/500: rewards 23.78 +/- 10.7
Iteration 105/500: rewards 19.22 +/- 8.65
Iteration 106/500: rewards 22.22 +/- 11.34
```

```
Iteration 107/500: rewards 20.53 +/- 9.56
Iteration 108/500: rewards 20.78 +/- 9.29
Iteration 109/500: rewards 17.22 +/- 5.89
Iteration 110/500: rewards 22.28 +/- 12.3
Iteration 111/500: rewards 17.84 +/- 7.81
Iteration 112/500: rewards 18.28 +/- 7.25
Iteration 113/500: rewards 19.81 +/- 9.53
Iteration 114/500: rewards 22.12 +/- 11.58
Iteration 115/500: rewards 20.38 +/- 9.14
Iteration 116/500: rewards 17.5 +/- 8.38
Iteration 117/500: rewards 21.03 +/- 11.38
Iteration 118/500: rewards 21.88 +/- 9.68
Iteration 119/500: rewards 19.69 +/- 10.99
Iteration 120/500: rewards 20.34 +/- 11.05
Iteration 121/500: rewards 20.06 +/- 7.89
Iteration 122/500: rewards 17.88 +/- 6.15
Iteration 123/500: rewards 23.75 +/- 14.73
Iteration 124/500: rewards 20.25 +/- 10.82
Iteration 125/500: rewards 20.88 +/- 9.22
Iteration 126/500: rewards 20.97 +/- 8.84
Iteration 127/500: rewards 20.09 +/- 8.29
Iteration 128/500: rewards 19.38 +/- 10.36
Iteration 129/500: rewards 19.66 +/- 11.98
Iteration 130/500: rewards 23.28 +/- 10.84
Iteration 131/500: rewards 19.25 +/- 7.63
Iteration 132/500: rewards 19.31 +/- 9.32
Iteration 133/500: rewards 17.06 +/- 5.23
Iteration 134/500: rewards 19.78 +/- 8.51
Iteration 135/500: rewards 21.16 +/- 13.83
Iteration 136/500: rewards 20.41 +/- 6.38
Iteration 137/500: rewards 19.12 +/- 8.97
Iteration 138/500: rewards 23.53 +/- 10.24
Iteration 139/500: rewards 24.16 +/- 20.1
Iteration 140/500: rewards 19.06 +/- 7.73
Iteration 141/500: rewards 23.69 +/- 11.16
Iteration 142/500: rewards 19.41 +/- 9.52
Iteration 143/500: rewards 22.78 +/- 15.66
Iteration 144/500: rewards 21.38 +/- 6.29
Iteration 145/500: rewards 20.62 +/- 9.82
Iteration 146/500: rewards 22.5 +/- 12.56
Iteration 147/500: rewards 21.03 +/- 8.16
Iteration 148/500: rewards 19.03 +/- 8.79
Iteration 149/500: rewards 18.06 +/- 6.85
Iteration 150/500: rewards 23.78 +/- 15.86
Iteration 151/500: rewards 19.34 +/- 10.39
Iteration 152/500: rewards 20.31 +/- 9.44
Iteration 153/500: rewards 24.56 +/- 12.32
Iteration 154/500: rewards 17.44 +/- 6.23
```

```
Iteration 155/500: rewards 18.62 +/- 9.62
Iteration 156/500: rewards 19.34 +/- 9.19
Iteration 157/500: rewards 21.47 +/- 11.12
Iteration 158/500: rewards 23.31 +/- 12.52
Iteration 159/500: rewards 21.09 +/- 8.66
Iteration 160/500: rewards 19.94 +/- 6.96
Iteration 161/500: rewards 24.28 +/- 11.71
Iteration 162/500: rewards 19.91 +/- 9.65
Iteration 163/500: rewards 18.34 +/- 8.72
Iteration 164/500: rewards 18.78 +/- 9.71
Iteration 165/500: rewards 20.94 +/- 8.61
Iteration 166/500: rewards 21.44 +/- 10.35
Iteration 167/500: rewards 20.22 +/- 8.77
Iteration 168/500: rewards 19.06 +/- 9.44
Iteration 169/500: rewards 19.28 +/- 6.31
Iteration 170/500: rewards 21.53 +/- 10.4
Iteration 171/500: rewards 21.59 +/- 12.75
Iteration 172/500: rewards 23.41 +/- 19.54
Iteration 173/500: rewards 22.56 +/- 14.17
Iteration 174/500: rewards 24.97 +/- 12.7
Iteration 175/500: rewards 22.41 +/- 10.13
Iteration 176/500: rewards 20.88 +/- 10.71
Iteration 177/500: rewards 21.47 +/- 13.07
Iteration 178/500: rewards 18.62 +/- 7.09
Iteration 179/500: rewards 20.5 +/- 8.0
Iteration 180/500: rewards 23.94 +/- 10.52
Iteration 181/500: rewards 19.62 +/- 11.09
Iteration 182/500: rewards 25.5 +/- 12.9
Iteration 183/500: rewards 24.75 +/- 14.4
Iteration 184/500: rewards 19.19 +/- 5.74
Iteration 185/500: rewards 19.53 +/- 9.88
Iteration 186/500: rewards 23.97 +/- 12.2
Iteration 187/500: rewards 23.12 +/- 11.12
Iteration 188/500: rewards 19.69 +/- 8.18
Iteration 189/500: rewards 23.75 +/- 14.03
Iteration 190/500: rewards 20.19 +/- 7.68
Iteration 191/500: rewards 23.75 +/- 13.86
Iteration 192/500: rewards 21.56 +/- 10.47
Iteration 193/500: rewards 23.28 +/- 16.14
Iteration 194/500: rewards 17.78 +/- 8.06
Iteration 195/500: rewards 20.22 +/- 8.07
Iteration 196/500: rewards 23.12 +/- 11.89
Iteration 197/500: rewards 22.19 +/- 9.04
Iteration 198/500: rewards 23.56 +/- 13.41
Iteration 199/500: rewards 18.94 +/- 7.62
Iteration 200/500: rewards 21.44 +/- 11.72
Iteration 201/500: rewards 19.56 +/- 10.75
Iteration 202/500: rewards 23.12 +/- 11.62
```

```
Iteration 203/500: rewards 21.62 +/- 9.86
Iteration 204/500: rewards 22.03 +/- 13.09
Iteration 205/500: rewards 20.66 +/- 9.0
Iteration 206/500: rewards 21.12 +/- 9.49
Iteration 207/500: rewards 21.16 +/- 15.54
Iteration 208/500: rewards 23.47 +/- 12.09
Iteration 209/500: rewards 20.81 +/- 12.52
Iteration 210/500: rewards 19.69 +/- 9.67
Iteration 211/500: rewards 26.22 +/- 16.26
Iteration 212/500: rewards 23.38 +/- 11.02
Iteration 213/500: rewards 22.69 +/- 13.92
Iteration 214/500: rewards 19.19 +/- 11.4
Iteration 215/500: rewards 20.31 +/- 11.81
Iteration 216/500: rewards 21.16 +/- 11.49
Iteration 217/500: rewards 23.66 +/- 13.33
Iteration 218/500: rewards 20.0 +/- 10.22
Iteration 219/500: rewards 20.62 +/- 14.67
Iteration 220/500: rewards 21.19 +/- 10.25
Iteration 221/500: rewards 23.5 +/- 9.18
Iteration 222/500: rewards 22.25 +/- 12.05
Iteration 223/500: rewards 20.22 +/- 8.21
Iteration 224/500: rewards 19.06 +/- 8.83
Iteration 225/500: rewards 19.28 +/- 7.69
Iteration 226/500: rewards 21.62 +/- 11.66
Iteration 227/500: rewards 20.78 +/- 9.28
Iteration 228/500: rewards 18.94 +/- 9.14
Iteration 229/500: rewards 24.19 +/- 12.56
Iteration 230/500: rewards 20.16 +/- 8.6
Iteration 231/500: rewards 20.12 +/- 8.33
Iteration 232/500: rewards 21.16 +/- 10.51
Iteration 233/500: rewards 18.97 +/- 8.94
Iteration 234/500: rewards 19.28 +/- 9.55
Iteration 235/500: rewards 20.38 +/- 8.44
Iteration 236/500: rewards 21.97 +/- 16.54
Iteration 237/500: rewards 18.56 +/- 7.68
Iteration 238/500: rewards 21.38 +/- 11.08
Iteration 239/500: rewards 26.03 +/- 13.12
Iteration 240/500: rewards 21.16 +/- 11.3
Iteration 241/500: rewards 22.03 +/- 11.78
Iteration 242/500: rewards 22.56 +/- 9.62
Iteration 243/500: rewards 22.09 +/- 10.58
Iteration 244/500: rewards 20.91 +/- 7.58
Iteration 245/500: rewards 20.53 +/- 12.81
Iteration 246/500: rewards 22.78 +/- 10.05
Iteration 247/500: rewards 23.16 +/- 10.95
Iteration 248/500: rewards 19.88 +/- 7.67
Iteration 249/500: rewards 20.03 +/- 10.42
Iteration 250/500: rewards 20.72 +/- 7.56
```

```
Iteration 251/500: rewards 23.06 +/- 11.64
Iteration 252/500: rewards 20.28 +/- 8.59
Iteration 253/500: rewards 18.25 +/- 6.35
Iteration 254/500: rewards 23.53 +/- 13.96
Iteration 255/500: rewards 20.44 +/- 12.04
Iteration 256/500: rewards 20.69 +/- 8.13
Iteration 257/500: rewards 24.72 +/- 12.65
Iteration 258/500: rewards 18.66 +/- 7.87
Iteration 259/500: rewards 17.56 +/- 7.96
Iteration 260/500: rewards 22.38 +/- 12.48
Iteration 261/500: rewards 20.69 +/- 11.47
Iteration 262/500: rewards 20.19 +/- 8.52
Iteration 263/500: rewards 20.09 +/- 11.18
Iteration 264/500: rewards 19.25 +/- 7.59
Iteration 265/500: rewards 24.91 +/- 12.61
Iteration 266/500: rewards 22.59 +/- 11.46
Iteration 267/500: rewards 22.56 +/- 11.73
Iteration 268/500: rewards 23.16 +/- 11.82
Iteration 269/500: rewards 23.12 +/- 9.95
Iteration 270/500: rewards 20.16 +/- 9.19
Iteration 271/500: rewards 20.59 +/- 15.78
Iteration 272/500: rewards 19.84 +/- 9.82
Iteration 273/500: rewards 21.06 +/- 12.93
Iteration 274/500: rewards 19.81 +/- 8.63
Iteration 275/500: rewards 21.78 +/- 14.58
Iteration 276/500: rewards 20.56 +/- 9.73
Iteration 277/500: rewards 21.5 +/- 9.3
Iteration 278/500: rewards 23.44 +/- 17.08
Iteration 279/500: rewards 26.0 +/- 21.98
Iteration 280/500: rewards 22.78 +/- 9.34
Iteration 281/500: rewards 24.12 +/- 8.94
Iteration 282/500: rewards 19.19 +/- 6.04
Iteration 283/500: rewards 19.28 +/- 10.05
Iteration 284/500: rewards 23.5 +/- 13.81
Iteration 285/500: rewards 18.12 +/- 6.33
Iteration 286/500: rewards 18.12 +/- 6.83
Iteration 287/500: rewards 18.47 +/- 9.7
Iteration 288/500: rewards 18.97 +/- 8.16
Iteration 289/500: rewards 22.69 +/- 13.74
Iteration 290/500: rewards 22.0 +/- 8.52
Iteration 291/500: rewards 24.28 +/- 16.8
Iteration 292/500: rewards 23.44 +/- 12.97
Iteration 293/500: rewards 22.06 +/- 11.2
Iteration 294/500: rewards 22.5 +/- 10.19
Iteration 295/500: rewards 17.94 +/- 10.31
Iteration 296/500: rewards 22.06 +/- 9.25
Iteration 297/500: rewards 19.12 +/- 11.99
Iteration 298/500: rewards 23.03 +/- 8.6
```

```
Iteration 299/500: rewards 17.91 +/- 10.03
Iteration 300/500: rewards 19.22 +/- 8.12
Iteration 301/500: rewards 20.19 +/- 9.65
Iteration 302/500: rewards 23.72 +/- 10.87
Iteration 303/500: rewards 23.81 +/- 15.24
Iteration 304/500: rewards 18.69 +/- 6.87
Iteration 305/500: rewards 19.59 +/- 8.43
Iteration 306/500: rewards 19.41 +/- 9.4
Iteration 307/500: rewards 19.0 +/- 6.76
Iteration 308/500: rewards 19.38 +/- 8.9
Iteration 309/500: rewards 20.88 +/- 9.75
Iteration 310/500: rewards 20.78 +/- 9.54
Iteration 311/500: rewards 19.97 +/- 10.58
Iteration 312/500: rewards 19.25 +/- 9.98
Iteration 313/500: rewards 22.88 +/- 12.0
Iteration 314/500: rewards 18.03 +/- 7.76
Iteration 315/500: rewards 26.44 +/- 15.92
Iteration 316/500: rewards 20.16 +/- 8.77
Iteration 317/500: rewards 18.78 +/- 7.61
Iteration 318/500: rewards 18.12 +/- 8.34
Iteration 319/500: rewards 21.91 +/- 8.21
Iteration 320/500: rewards 22.84 +/- 13.18
Iteration 321/500: rewards 22.41 +/- 12.46
Iteration 322/500: rewards 21.0 +/- 9.95
Iteration 323/500: rewards 21.88 +/- 9.5
Iteration 324/500: rewards 20.25 +/- 9.37
Iteration 325/500: rewards 20.38 +/- 10.46
Iteration 326/500: rewards 19.97 +/- 9.98
Iteration 327/500: rewards 21.59 +/- 12.45
Iteration 328/500: rewards 20.41 +/- 9.63
Iteration 329/500: rewards 20.06 +/- 8.63
Iteration 330/500: rewards 22.28 +/- 12.69
Iteration 331/500: rewards 20.78 +/- 11.99
Iteration 332/500: rewards 21.91 +/- 14.6
Iteration 333/500: rewards 23.84 +/- 12.33
Iteration 334/500: rewards 19.78 +/- 9.28
Iteration 335/500: rewards 20.38 +/- 10.77
Iteration 336/500: rewards 20.75 +/- 8.45
Iteration 337/500: rewards 19.91 +/- 11.59
Iteration 338/500: rewards 23.47 +/- 12.5
Iteration 339/500: rewards 21.75 +/- 8.36
Iteration 340/500: rewards 19.25 +/- 8.2
Iteration 341/500: rewards 20.28 +/- 9.69
Iteration 342/500: rewards 23.16 +/- 12.84
Iteration 343/500: rewards 20.16 +/- 8.45
Iteration 344/500: rewards 18.31 +/- 9.22
Iteration 345/500: rewards 19.59 +/- 10.42
Iteration 346/500: rewards 17.44 +/- 6.6
```

```
Iteration 347/500: rewards 17.69 +/- 7.51
Iteration 348/500: rewards 25.06 +/- 12.09
Iteration 349/500: rewards 21.72 +/- 14.17
Iteration 350/500: rewards 22.97 +/- 10.3
Iteration 351/500: rewards 19.91 +/- 9.73
Iteration 352/500: rewards 20.16 +/- 11.8
Iteration 353/500: rewards 19.88 +/- 9.37
Iteration 354/500: rewards 21.62 +/- 10.09
Iteration 355/500: rewards 19.19 +/- 8.67
Iteration 356/500: rewards 18.44 +/- 7.98
Iteration 357/500: rewards 19.62 +/- 9.77
Iteration 358/500: rewards 21.22 +/- 8.42
Iteration 359/500: rewards 18.78 +/- 9.3
Iteration 360/500: rewards 22.75 +/- 13.62
Iteration 361/500: rewards 22.44 +/- 11.9
Iteration 362/500: rewards 17.28 +/- 7.95
Iteration 363/500: rewards 20.5 +/- 9.91
Iteration 364/500: rewards 20.41 +/- 12.31
Iteration 365/500: rewards 20.91 +/- 13.28
Iteration 366/500: rewards 19.44 +/- 8.66
Iteration 367/500: rewards 20.44 +/- 8.33
Iteration 368/500: rewards 21.19 +/- 13.01
Iteration 369/500: rewards 17.56 +/- 6.26
Iteration 370/500: rewards 19.09 +/- 8.35
Iteration 371/500: rewards 20.88 +/- 7.8
Iteration 372/500: rewards 21.47 +/- 8.66
Iteration 373/500: rewards 17.62 +/- 9.01
Iteration 374/500: rewards 19.5 +/- 7.36
Iteration 375/500: rewards 18.94 +/- 7.77
Iteration 376/500: rewards 20.06 +/- 11.47
Iteration 377/500: rewards 18.34 +/- 8.08
Iteration 378/500: rewards 18.69 +/- 9.16
Iteration 379/500: rewards 18.25 +/- 6.86
Iteration 380/500: rewards 19.5 +/- 7.55
Iteration 381/500: rewards 22.22 +/- 7.96
Iteration 382/500: rewards 22.69 +/- 14.65
Iteration 383/500: rewards 17.0 +/- 9.25
Iteration 384/500: rewards 18.44 +/- 7.28
Iteration 385/500: rewards 17.44 +/- 6.62
Iteration 386/500: rewards 18.25 +/- 9.06
Iteration 387/500: rewards 20.38 +/- 8.79
Iteration 388/500: rewards 21.75 +/- 14.67
Iteration 389/500: rewards 19.88 +/- 8.76
Iteration 390/500: rewards 18.06 +/- 7.66
Iteration 391/500: rewards 19.69 +/- 6.96
Iteration 392/500: rewards 17.34 +/- 7.01
Iteration 393/500: rewards 26.41 +/- 17.44
Iteration 394/500: rewards 19.78 +/- 9.34
```

```
Iteration 395/500: rewards 21.19 +/- 10.0
Iteration 396/500: rewards 19.31 +/- 7.67
Iteration 397/500: rewards 19.84 +/- 9.0
Iteration 398/500: rewards 21.06 +/- 10.63
Iteration 399/500: rewards 19.16 +/- 10.09
Iteration 400/500: rewards 17.5 +/- 5.89
Iteration 401/500: rewards 18.69 +/- 8.84
Iteration 402/500: rewards 18.69 +/- 7.7
Iteration 403/500: rewards 20.47 +/- 11.7
Iteration 404/500: rewards 16.31 +/- 5.5
Iteration 405/500: rewards 23.06 +/- 11.62
Iteration 406/500: rewards 19.66 +/- 11.58
Iteration 407/500: rewards 20.06 +/- 8.62
Iteration 408/500: rewards 21.31 +/- 11.5
Iteration 409/500: rewards 24.0 +/- 11.74
Iteration 410/500: rewards 21.5 +/- 11.62
Iteration 411/500: rewards 19.16 +/- 10.28
Iteration 412/500: rewards 17.72 +/- 9.62
Iteration 413/500: rewards 22.09 +/- 11.38
Iteration 414/500: rewards 20.22 +/- 10.18
Iteration 415/500: rewards 18.56 +/- 6.57
Iteration 416/500: rewards 20.31 +/- 6.99
Iteration 417/500: rewards 20.31 +/- 11.49
Iteration 418/500: rewards 22.25 +/- 12.42
Iteration 419/500: rewards 19.81 +/- 9.82
Iteration 420/500: rewards 19.34 +/- 7.04
Iteration 421/500: rewards 22.47 +/- 12.31
Iteration 422/500: rewards 18.56 +/- 8.44
Iteration 423/500: rewards 20.84 +/- 10.07
Iteration 424/500: rewards 18.59 +/- 8.85
Iteration 425/500: rewards 19.88 +/- 11.19
Iteration 426/500: rewards 23.72 +/- 14.5
Iteration 427/500: rewards 19.25 +/- 9.23
Iteration 428/500: rewards 20.03 +/- 10.63
Iteration 429/500: rewards 20.12 +/- 9.1
Iteration 430/500: rewards 17.53 +/- 6.08
Iteration 431/500: rewards 17.81 +/- 8.78
Iteration 432/500: rewards 19.5 +/- 10.19
Iteration 433/500: rewards 20.84 +/- 11.48
Iteration 434/500: rewards 18.78 +/- 9.52
Iteration 435/500: rewards 19.81 +/- 13.39
Iteration 436/500: rewards 17.31 +/- 5.76
Iteration 437/500: rewards 19.81 +/- 11.2
Iteration 438/500: rewards 19.03 +/- 9.03
Iteration 439/500: rewards 18.0 +/- 7.4
Iteration 440/500: rewards 20.38 +/- 8.73
Iteration 441/500: rewards 19.38 +/- 9.02
Iteration 442/500: rewards 20.69 +/- 11.46
```

```
Iteration 443/500: rewards 19.62 +/- 9.33
Iteration 444/500: rewards 18.84 +/- 9.3
Iteration 445/500: rewards 18.81 +/- 8.74
Iteration 446/500: rewards 20.0 +/- 15.03
Iteration 447/500: rewards 19.78 +/- 8.75
Iteration 448/500: rewards 21.25 +/- 10.26
Iteration 449/500: rewards 15.62 +/- 5.14
Iteration 450/500: rewards 20.34 +/- 8.59
Iteration 451/500: rewards 21.38 +/- 11.19
Iteration 452/500: rewards 17.0 +/- 7.19
Iteration 453/500: rewards 16.09 +/- 6.0
Iteration 454/500: rewards 15.41 +/- 5.99
Iteration 455/500: rewards 17.56 +/- 7.18
Iteration 456/500: rewards 18.78 +/- 7.02
Iteration 457/500: rewards 18.78 +/- 9.97
Iteration 458/500: rewards 18.41 +/- 9.48
Iteration 459/500: rewards 18.81 +/- 8.5
Iteration 460/500: rewards 18.12 +/- 7.2
Iteration 461/500: rewards 19.31 +/- 7.75
Iteration 462/500: rewards 19.5 +/- 8.84
Iteration 463/500: rewards 19.47 +/- 7.64
Iteration 464/500: rewards 20.75 +/- 11.55
Iteration 465/500: rewards 18.28 +/- 8.52
Iteration 466/500: rewards 20.72 +/- 8.97
Iteration 467/500: rewards 18.94 +/- 7.83
Iteration 468/500: rewards 19.75 +/- 8.68
Iteration 469/500: rewards 19.47 +/- 9.12
Iteration 470/500: rewards 19.09 +/- 9.36
Iteration 471/500: rewards 17.62 +/- 8.75
Iteration 472/500: rewards 21.38 +/- 10.73
Iteration 473/500: rewards 22.5 +/- 14.53
Iteration 474/500: rewards 18.16 +/- 7.2
Iteration 475/500: rewards 21.12 +/- 10.45
Iteration 476/500: rewards 18.44 +/- 4.93
Iteration 477/500: rewards 19.03 +/- 6.65
Iteration 478/500: rewards 18.91 +/- 7.19
Iteration 479/500: rewards 18.03 +/- 8.26
Iteration 480/500: rewards 17.34 +/- 6.28
Iteration 481/500: rewards 17.12 +/- 6.0
Iteration 482/500: rewards 19.03 +/- 12.05
Iteration 483/500: rewards 21.56 +/- 12.87
Iteration 484/500: rewards 16.47 +/- 5.94
Iteration 485/500: rewards 16.78 +/- 7.58
Iteration 486/500: rewards 16.66 +/- 6.65
Iteration 487/500: rewards 22.0 +/- 12.74
Iteration 488/500: rewards 16.47 +/- 5.74
Iteration 489/500: rewards 18.28 +/- 7.58
Iteration 490/500: rewards 20.53 +/- 8.63
```

```
Iteration 491/500: rewards 18.41 +/- 6.47
Iteration 492/500: rewards 18.69 +/- 8.65
Iteration 493/500: rewards 20.25 +/- 13.38
Iteration 494/500: rewards 18.72 +/- 12.45
Iteration 495/500: rewards 17.81 +/- 7.43
Iteration 496/500: rewards 19.94 +/- 10.34
Iteration 497/500: rewards 18.03 +/- 8.24
Iteration 498/500: rewards 19.88 +/- 7.91
Iteration 499/500: rewards 19.25 +/- 11.03
Iteration 500/500: rewards 17.69 +/- 8.43
The average reward is 18.834375
the device is: cpu
------------------------------------------------
The gamma chosen is: 0.95
The value lr chosen is 0.01
The policy lr chosen is 0.0001
Iteration 1/500: rewards 19.0 +/- 7.26
Iteration 2/500: rewards 18.59 +/- 9.64
Iteration 3/500: rewards 17.69 +/- 8.52
Iteration 4/500: rewards 20.84 +/- 11.63
Iteration 5/500: rewards 17.03 +/- 8.93
Iteration 6/500: rewards 19.59 +/- 9.46
Iteration 7/500: rewards 18.47 +/- 10.26
Iteration 8/500: rewards 18.94 +/- 7.55
Iteration 9/500: rewards 21.56 +/- 7.56
Iteration 10/500: rewards 21.53 +/- 11.05
Iteration 11/500: rewards 17.38 +/- 5.35
Iteration 12/500: rewards 23.78 +/- 9.79
Iteration 13/500: rewards 19.84 +/- 10.66
Iteration 14/500: rewards 18.53 +/- 10.12
Iteration 15/500: rewards 21.5 +/- 13.1
Iteration 16/500: rewards 20.03 +/- 8.54
Iteration 17/500: rewards 17.25 +/- 8.14
Iteration 18/500: rewards 19.53 +/- 7.96
Iteration 19/500: rewards 18.03 +/- 6.71
Iteration 20/500: rewards 18.25 +/- 8.32
Iteration 21/500: rewards 17.41 +/- 4.88
Iteration 22/500: rewards 24.31 +/- 15.47
Iteration 23/500: rewards 17.59 +/- 7.38
Iteration 24/500: rewards 18.25 +/- 7.76
Iteration 25/500: rewards 18.59 +/- 10.15
Iteration 26/500: rewards 19.03 +/- 10.81
Iteration 27/500: rewards 18.81 +/- 8.36
Iteration 28/500: rewards 19.47 +/- 8.77
Iteration 29/500: rewards 20.19 +/- 9.18
Iteration 30/500: rewards 19.44 +/- 9.91
Iteration 31/500: rewards 18.31 +/- 7.84
Iteration 32/500: rewards 18.81 +/- 10.83
```

```
Iteration 33/500: rewards 19.31 +/- 8.52
Iteration 34/500: rewards 18.62 +/- 7.59
Iteration 35/500: rewards 19.44 +/- 7.86
Iteration 36/500: rewards 20.88 +/- 12.21
Iteration 37/500: rewards 16.47 +/- 6.36
Iteration 38/500: rewards 20.75 +/- 9.1
Iteration 39/500: rewards 14.94 +/- 5.88
Iteration 40/500: rewards 18.09 +/- 6.33
Iteration 41/500: rewards 18.53 +/- 8.5
Iteration 42/500: rewards 20.38 +/- 11.29
Iteration 43/500: rewards 18.12 +/- 10.9
Iteration 44/500: rewards 19.75 +/- 9.15
Iteration 45/500: rewards 22.12 +/- 11.21
Iteration 46/500: rewards 19.75 +/- 10.15
Iteration 47/500: rewards 21.91 +/- 11.95
Iteration 48/500: rewards 19.53 +/- 10.65
Iteration 49/500: rewards 19.91 +/- 9.07
Iteration 50/500: rewards 19.03 +/- 9.47
Iteration 51/500: rewards 25.44 +/- 14.21
Iteration 52/500: rewards 18.66 +/- 8.66
Iteration 53/500: rewards 19.97 +/- 9.82
Iteration 54/500: rewards 13.88 +/- 2.92
Iteration 55/500: rewards 22.75 +/- 15.64
Iteration 56/500: rewards 19.84 +/- 9.89
Iteration 57/500: rewards 20.25 +/- 10.11
Iteration 58/500: rewards 17.84 +/- 6.15
Iteration 59/500: rewards 17.66 +/- 8.31
Iteration 60/500: rewards 19.97 +/- 11.78
Iteration 61/500: rewards 18.12 +/- 7.66
Iteration 62/500: rewards 20.56 +/- 10.74
Iteration 63/500: rewards 20.03 +/- 10.21
Iteration 64/500: rewards 19.12 +/- 6.95
Iteration 65/500: rewards 19.16 +/- 8.62
Iteration 66/500: rewards 19.81 +/- 9.38
Iteration 67/500: rewards 23.97 +/- 14.89
Iteration 68/500: rewards 19.34 +/- 10.66
Iteration 69/500: rewards 19.97 +/- 8.68
Iteration 70/500: rewards 17.78 +/- 7.73
Iteration 71/500: rewards 21.84 +/- 9.98
Iteration 72/500: rewards 17.97 +/- 6.0
Iteration 73/500: rewards 21.16 +/- 12.43
Iteration 74/500: rewards 19.38 +/- 10.38
Iteration 75/500: rewards 21.44 +/- 13.81
Iteration 76/500: rewards 20.62 +/- 12.83
Iteration 77/500: rewards 18.88 +/- 10.91
Iteration 78/500: rewards 22.88 +/- 11.32
Iteration 79/500: rewards 24.03 +/- 16.22
Iteration 80/500: rewards 20.66 +/- 10.71
```

```
Iteration 81/500: rewards 18.84 +/- 10.6
Iteration 82/500: rewards 21.72 +/- 14.75
Iteration 83/500: rewards 16.66 +/- 5.6
Iteration 84/500: rewards 18.03 +/- 7.89
Iteration 85/500: rewards 19.25 +/- 8.58
Iteration 86/500: rewards 20.31 +/- 11.48
Iteration 87/500: rewards 18.75 +/- 6.71
Iteration 88/500: rewards 19.19 +/- 10.1
Iteration 89/500: rewards 22.66 +/- 11.07
Iteration 90/500: rewards 21.16 +/- 10.33
Iteration 91/500: rewards 20.5 +/- 6.82
Iteration 92/500: rewards 18.94 +/- 11.37
Iteration 93/500: rewards 19.72 +/- 8.99
Iteration 94/500: rewards 22.94 +/- 10.97
Iteration 95/500: rewards 19.38 +/- 7.8
Iteration 96/500: rewards 17.75 +/- 7.88
Iteration 97/500: rewards 21.69 +/- 8.76
Iteration 98/500: rewards 21.38 +/- 14.12
Iteration 99/500: rewards 20.41 +/- 11.17
Iteration 100/500: rewards 18.75 +/- 7.23
Iteration 101/500: rewards 19.12 +/- 7.41
Iteration 102/500: rewards 19.53 +/- 10.51
Iteration 103/500: rewards 24.31 +/- 14.55
Iteration 104/500: rewards 19.81 +/- 7.94
Iteration 105/500: rewards 19.62 +/- 12.23
Iteration 106/500: rewards 19.25 +/- 7.98
Iteration 107/500: rewards 17.91 +/- 9.21
Iteration 108/500: rewards 17.16 +/- 6.24
Iteration 109/500: rewards 21.03 +/- 10.32
Iteration 110/500: rewards 19.66 +/- 9.18
Iteration 111/500: rewards 19.03 +/- 10.46
Iteration 112/500: rewards 15.97 +/- 5.48
Iteration 113/500: rewards 18.47 +/- 8.06
Iteration 114/500: rewards 20.38 +/- 7.97
Iteration 115/500: rewards 21.62 +/- 12.68
Iteration 116/500: rewards 18.81 +/- 6.38
Iteration 117/500: rewards 19.16 +/- 10.8
Iteration 118/500: rewards 19.47 +/- 8.46
Iteration 119/500: rewards 20.34 +/- 8.56
Iteration 120/500: rewards 18.53 +/- 7.59
Iteration 121/500: rewards 19.09 +/- 6.72
Iteration 122/500: rewards 22.31 +/- 10.09
Iteration 123/500: rewards 21.91 +/- 11.91
Iteration 124/500: rewards 22.81 +/- 11.91
Iteration 125/500: rewards 18.03 +/- 9.15
Iteration 126/500: rewards 20.34 +/- 10.21
Iteration 127/500: rewards 22.91 +/- 11.86
Iteration 128/500: rewards 18.41 +/- 9.04
```

```
Iteration 129/500: rewards 18.78 +/- 8.53
Iteration 130/500: rewards 19.5 +/- 8.38
Iteration 131/500: rewards 20.88 +/- 11.77
Iteration 132/500: rewards 22.72 +/- 18.31
Iteration 133/500: rewards 21.03 +/- 11.0
Iteration 134/500: rewards 21.84 +/- 14.19
Iteration 135/500: rewards 24.56 +/- 9.22
Iteration 136/500: rewards 21.91 +/- 12.36
Iteration 137/500: rewards 19.19 +/- 7.97
Iteration 138/500: rewards 19.0 +/- 8.58
Iteration 139/500: rewards 22.0 +/- 11.54
Iteration 140/500: rewards 18.78 +/- 8.95
Iteration 141/500: rewards 17.88 +/- 6.81
Iteration 142/500: rewards 19.5 +/- 7.5
Iteration 143/500: rewards 22.03 +/- 11.14
Iteration 144/500: rewards 20.78 +/- 9.87
Iteration 145/500: rewards 19.09 +/- 5.64
Iteration 146/500: rewards 21.47 +/- 14.93
Iteration 147/500: rewards 20.56 +/- 9.12
Iteration 148/500: rewards 20.75 +/- 10.56
Iteration 149/500: rewards 16.59 +/- 6.39
Iteration 150/500: rewards 20.69 +/- 11.73
Iteration 151/500: rewards 19.5 +/- 8.58
Iteration 152/500: rewards 20.91 +/- 10.11
Iteration 153/500: rewards 21.31 +/- 11.49
Iteration 154/500: rewards 21.53 +/- 10.54
Iteration 155/500: rewards 18.81 +/- 8.94
Iteration 156/500: rewards 24.28 +/- 17.78
Iteration 157/500: rewards 19.12 +/- 9.76
Iteration 158/500: rewards 21.41 +/- 12.61
Iteration 159/500: rewards 24.31 +/- 9.43
Iteration 160/500: rewards 19.19 +/- 9.24
Iteration 161/500: rewards 20.56 +/- 11.13
Iteration 162/500: rewards 20.69 +/- 8.85
Iteration 163/500: rewards 21.34 +/- 11.6
Iteration 164/500: rewards 21.31 +/- 15.2
Iteration 165/500: rewards 21.66 +/- 10.95
Iteration 166/500: rewards 18.94 +/- 12.77
Iteration 167/500: rewards 20.38 +/- 9.26
Iteration 168/500: rewards 19.19 +/- 8.06
Iteration 169/500: rewards 23.91 +/- 17.29
Iteration 170/500: rewards 20.31 +/- 12.57
Iteration 171/500: rewards 20.88 +/- 8.62
Iteration 172/500: rewards 18.28 +/- 7.97
Iteration 173/500: rewards 23.75 +/- 11.28
Iteration 174/500: rewards 23.69 +/- 11.28
Iteration 175/500: rewards 20.94 +/- 10.44
Iteration 176/500: rewards 20.53 +/- 10.25
```

```
Iteration 177/500: rewards 19.5 +/- 8.72
Iteration 178/500: rewards 20.25 +/- 13.73
Iteration 179/500: rewards 21.09 +/- 18.48
Iteration 180/500: rewards 23.38 +/- 13.55
Iteration 181/500: rewards 24.59 +/- 15.49
Iteration 182/500: rewards 21.22 +/- 8.75
Iteration 183/500: rewards 23.88 +/- 12.15
Iteration 184/500: rewards 22.88 +/- 10.38
Iteration 185/500: rewards 18.72 +/- 7.56
Iteration 186/500: rewards 17.62 +/- 7.23
Iteration 187/500: rewards 21.19 +/- 14.74
Iteration 188/500: rewards 18.97 +/- 7.21
Iteration 189/500: rewards 22.44 +/- 11.27
Iteration 190/500: rewards 20.97 +/- 10.31
Iteration 191/500: rewards 20.72 +/- 12.39
Iteration 192/500: rewards 18.56 +/- 8.45
Iteration 193/500: rewards 22.0 +/- 10.98
Iteration 194/500: rewards 19.69 +/- 8.86
Iteration 195/500: rewards 23.31 +/- 12.93
Iteration 196/500: rewards 22.62 +/- 12.21
Iteration 197/500: rewards 18.31 +/- 8.53
Iteration 198/500: rewards 25.38 +/- 11.94
Iteration 199/500: rewards 24.97 +/- 13.09
Iteration 200/500: rewards 19.72 +/- 8.3
Iteration 201/500: rewards 21.59 +/- 17.21
Iteration 202/500: rewards 21.47 +/- 10.88
Iteration 203/500: rewards 22.0 +/- 9.88
Iteration 204/500: rewards 16.34 +/- 5.82
Iteration 205/500: rewards 22.03 +/- 11.41
Iteration 206/500: rewards 18.78 +/- 7.25
Iteration 207/500: rewards 23.5 +/- 11.24
Iteration 208/500: rewards 21.41 +/- 9.73
Iteration 209/500: rewards 22.88 +/- 13.61
Iteration 210/500: rewards 22.81 +/- 12.67
Iteration 211/500: rewards 18.09 +/- 7.02
Iteration 212/500: rewards 16.78 +/- 6.9
Iteration 213/500: rewards 24.56 +/- 13.8
Iteration 214/500: rewards 18.81 +/- 7.6
Iteration 215/500: rewards 25.84 +/- 13.83
Iteration 216/500: rewards 23.19 +/- 13.91
Iteration 217/500: rewards 18.78 +/- 7.44
Iteration 218/500: rewards 19.66 +/- 11.65
Iteration 219/500: rewards 22.12 +/- 11.61
Iteration 220/500: rewards 21.12 +/- 17.49
Iteration 221/500: rewards 18.66 +/- 8.74
Iteration 222/500: rewards 22.47 +/- 9.79
Iteration 223/500: rewards 21.03 +/- 11.07
Iteration 224/500: rewards 19.41 +/- 8.9
```

```
Iteration 225/500: rewards 21.19 +/- 12.49
Iteration 226/500: rewards 19.28 +/- 7.51
Iteration 227/500: rewards 21.88 +/- 11.74
Iteration 228/500: rewards 18.75 +/- 9.0
Iteration 229/500: rewards 21.16 +/- 9.94
Iteration 230/500: rewards 19.47 +/- 8.5
Iteration 231/500: rewards 20.38 +/- 10.25
Iteration 232/500: rewards 20.41 +/- 8.26
Iteration 233/500: rewards 21.91 +/- 11.18
Iteration 234/500: rewards 18.78 +/- 7.57
Iteration 235/500: rewards 18.09 +/- 9.86
Iteration 236/500: rewards 18.78 +/- 7.74
Iteration 237/500: rewards 21.12 +/- 9.6
Iteration 238/500: rewards 22.06 +/- 12.04
Iteration 239/500: rewards 17.88 +/- 5.56
Iteration 240/500: rewards 22.72 +/- 10.09
Iteration 241/500: rewards 21.5 +/- 11.17
Iteration 242/500: rewards 21.62 +/- 11.2
Iteration 243/500: rewards 20.53 +/- 10.68
Iteration 244/500: rewards 24.44 +/- 14.11
Iteration 245/500: rewards 23.56 +/- 13.44
Iteration 246/500: rewards 24.66 +/- 12.15
Iteration 247/500: rewards 19.44 +/- 7.32
Iteration 248/500: rewards 19.88 +/- 8.85
Iteration 249/500: rewards 20.03 +/- 9.84
Iteration 250/500: rewards 19.47 +/- 7.8
Iteration 251/500: rewards 22.09 +/- 9.69
Iteration 252/500: rewards 21.0 +/- 10.46
Iteration 253/500: rewards 24.0 +/- 14.87
Iteration 254/500: rewards 21.44 +/- 9.49
Iteration 255/500: rewards 19.06 +/- 7.78
Iteration 256/500: rewards 22.25 +/- 12.13
Iteration 257/500: rewards 19.06 +/- 7.83
Iteration 258/500: rewards 21.34 +/- 8.8
Iteration 259/500: rewards 18.91 +/- 8.39
Iteration 260/500: rewards 19.0 +/- 9.45
Iteration 261/500: rewards 21.0 +/- 11.76
Iteration 262/500: rewards 22.59 +/- 11.86
Iteration 263/500: rewards 22.03 +/- 13.66
Iteration 264/500: rewards 22.12 +/- 11.2
Iteration 265/500: rewards 19.69 +/- 9.16
Iteration 266/500: rewards 23.84 +/- 11.75
Iteration 267/500: rewards 22.25 +/- 14.52
Iteration 268/500: rewards 23.34 +/- 13.84
Iteration 269/500: rewards 22.22 +/- 14.68
Iteration 270/500: rewards 19.5 +/- 9.67
Iteration 271/500: rewards 22.78 +/- 10.14
Iteration 272/500: rewards 22.19 +/- 12.1
```

```
Iteration 273/500: rewards 22.22 +/- 13.03
Iteration 274/500: rewards 19.69 +/- 9.95
Iteration 275/500: rewards 22.16 +/- 11.42
Iteration 276/500: rewards 20.41 +/- 13.41
Iteration 277/500: rewards 22.03 +/- 12.1
Iteration 278/500: rewards 25.69 +/- 16.99
Iteration 279/500: rewards 24.03 +/- 12.11
Iteration 280/500: rewards 21.47 +/- 10.9
Iteration 281/500: rewards 22.28 +/- 10.08
Iteration 282/500: rewards 21.38 +/- 9.77
Iteration 283/500: rewards 22.59 +/- 11.33
Iteration 284/500: rewards 22.66 +/- 14.16
Iteration 285/500: rewards 20.97 +/- 8.93
Iteration 286/500: rewards 25.56 +/- 14.61
Iteration 287/500: rewards 23.0 +/- 15.43
Iteration 288/500: rewards 23.12 +/- 10.51
Iteration 289/500: rewards 18.25 +/- 6.79
Iteration 290/500: rewards 21.34 +/- 9.57
Iteration 291/500: rewards 19.78 +/- 8.92
Iteration 292/500: rewards 20.28 +/- 9.19
Iteration 293/500: rewards 19.72 +/- 10.86
Iteration 294/500: rewards 23.69 +/- 14.23
Iteration 295/500: rewards 22.41 +/- 9.01
Iteration 296/500: rewards 24.5 +/- 12.45
Iteration 297/500: rewards 21.69 +/- 10.38
Iteration 298/500: rewards 20.88 +/- 8.93
Iteration 299/500: rewards 22.59 +/- 12.59
Iteration 300/500: rewards 18.62 +/- 8.4
Iteration 301/500: rewards 18.16 +/- 10.29
Iteration 302/500: rewards 19.88 +/- 9.72
Iteration 303/500: rewards 22.16 +/- 12.92
Iteration 304/500: rewards 21.72 +/- 10.42
Iteration 305/500: rewards 18.94 +/- 8.01
Iteration 306/500: rewards 23.56 +/- 12.09
Iteration 307/500: rewards 22.38 +/- 10.16
Iteration 308/500: rewards 19.53 +/- 9.4
Iteration 309/500: rewards 19.59 +/- 9.65
Iteration 310/500: rewards 21.84 +/- 9.25
Iteration 311/500: rewards 21.97 +/- 7.86
Iteration 312/500: rewards 21.22 +/- 10.62
Iteration 313/500: rewards 19.56 +/- 8.79
Iteration 314/500: rewards 25.78 +/- 20.31
Iteration 315/500: rewards 21.62 +/- 10.45
Iteration 316/500: rewards 20.84 +/- 6.76
Iteration 317/500: rewards 25.03 +/- 15.46
Iteration 318/500: rewards 20.59 +/- 10.02
Iteration 319/500: rewards 24.84 +/- 12.32
Iteration 320/500: rewards 19.66 +/- 8.05
```

```
Iteration 321/500: rewards 20.91 +/- 10.24
Iteration 322/500: rewards 18.94 +/- 8.66
Iteration 323/500: rewards 21.62 +/- 13.67
Iteration 324/500: rewards 21.94 +/- 7.65
Iteration 325/500: rewards 22.41 +/- 11.67
Iteration 326/500: rewards 20.03 +/- 8.74
Iteration 327/500: rewards 22.25 +/- 10.21
Iteration 328/500: rewards 22.5 +/- 13.35
Iteration 329/500: rewards 18.22 +/- 10.69
Iteration 330/500: rewards 20.09 +/- 8.06
Iteration 331/500: rewards 18.38 +/- 7.26
Iteration 332/500: rewards 21.78 +/- 9.46
Iteration 333/500: rewards 20.75 +/- 10.85
Iteration 334/500: rewards 20.56 +/- 9.47
Iteration 335/500: rewards 20.12 +/- 8.76
Iteration 336/500: rewards 23.78 +/- 10.76
Iteration 337/500: rewards 23.28 +/- 13.44
Iteration 338/500: rewards 21.16 +/- 9.27
Iteration 339/500: rewards 20.0 +/- 9.6
Iteration 340/500: rewards 20.53 +/- 10.05
Iteration 341/500: rewards 22.38 +/- 14.97
Iteration 342/500: rewards 24.31 +/- 12.23
Iteration 343/500: rewards 24.47 +/- 16.76
Iteration 344/500: rewards 24.31 +/- 13.6
Iteration 345/500: rewards 17.84 +/- 7.97
Iteration 346/500: rewards 22.94 +/- 8.69
Iteration 347/500: rewards 18.84 +/- 6.87
Iteration 348/500: rewards 21.16 +/- 9.58
Iteration 349/500: rewards 21.44 +/- 9.22
Iteration 350/500: rewards 20.84 +/- 12.49
Iteration 351/500: rewards 20.5 +/- 8.03
Iteration 352/500: rewards 22.94 +/- 10.23
Iteration 353/500: rewards 23.16 +/- 9.91
Iteration 354/500: rewards 21.78 +/- 14.73
Iteration 355/500: rewards 21.38 +/- 11.77
Iteration 356/500: rewards 22.41 +/- 12.9
Iteration 357/500: rewards 22.53 +/- 13.17
Iteration 358/500: rewards 18.94 +/- 6.86
Iteration 359/500: rewards 20.47 +/- 10.84
Iteration 360/500: rewards 21.47 +/- 9.0
Iteration 361/500: rewards 22.75 +/- 10.89
Iteration 362/500: rewards 21.91 +/- 10.68
Iteration 363/500: rewards 21.78 +/- 8.75
Iteration 364/500: rewards 22.56 +/- 10.56
Iteration 365/500: rewards 19.91 +/- 7.85
Iteration 366/500: rewards 24.22 +/- 11.53
Iteration 367/500: rewards 20.09 +/- 9.38
Iteration 368/500: rewards 21.62 +/- 8.59
```

```
Iteration 369/500: rewards 23.94 +/- 15.94
Iteration 370/500: rewards 22.94 +/- 9.93
Iteration 371/500: rewards 20.94 +/- 9.75
Iteration 372/500: rewards 22.19 +/- 9.76
Iteration 373/500: rewards 18.97 +/- 13.58
Iteration 374/500: rewards 20.75 +/- 11.99
Iteration 375/500: rewards 21.09 +/- 10.44
Iteration 376/500: rewards 25.0 +/- 15.94
Iteration 377/500: rewards 24.78 +/- 10.58
Iteration 378/500: rewards 23.94 +/- 11.54
Iteration 379/500: rewards 23.47 +/- 13.63
Iteration 380/500: rewards 19.53 +/- 9.95
Iteration 381/500: rewards 17.22 +/- 6.59
Iteration 382/500: rewards 18.69 +/- 7.78
Iteration 383/500: rewards 27.84 +/- 14.93
Iteration 384/500: rewards 23.94 +/- 12.44
Iteration 385/500: rewards 24.66 +/- 10.32
Iteration 386/500: rewards 24.25 +/- 15.79
Iteration 387/500: rewards 19.84 +/- 9.06
Iteration 388/500: rewards 24.59 +/- 15.0
Iteration 389/500: rewards 23.91 +/- 13.88
Iteration 390/500: rewards 20.28 +/- 11.84
Iteration 391/500: rewards 21.25 +/- 9.0
Iteration 392/500: rewards 21.06 +/- 10.4
Iteration 393/500: rewards 21.34 +/- 15.71
Iteration 394/500: rewards 21.88 +/- 10.39
Iteration 395/500: rewards 23.62 +/- 13.08
Iteration 396/500: rewards 23.22 +/- 11.83
Iteration 397/500: rewards 20.97 +/- 9.03
Iteration 398/500: rewards 25.0 +/- 13.67
Iteration 399/500: rewards 20.72 +/- 9.83
Iteration 400/500: rewards 20.91 +/- 8.76
Iteration 401/500: rewards 21.69 +/- 7.89
Iteration 402/500: rewards 23.31 +/- 11.01
Iteration 403/500: rewards 23.25 +/- 14.36
Iteration 404/500: rewards 24.97 +/- 20.87
Iteration 405/500: rewards 19.75 +/- 9.39
Iteration 406/500: rewards 22.75 +/- 10.37
Iteration 407/500: rewards 28.5 +/- 18.03
Iteration 408/500: rewards 22.53 +/- 12.78
Iteration 409/500: rewards 23.66 +/- 12.93
Iteration 410/500: rewards 22.03 +/- 12.54
Iteration 411/500: rewards 20.41 +/- 6.41
Iteration 412/500: rewards 22.12 +/- 12.2
Iteration 413/500: rewards 23.47 +/- 13.35
Iteration 414/500: rewards 23.72 +/- 12.8
Iteration 415/500: rewards 21.91 +/- 10.66
Iteration 416/500: rewards 23.0 +/- 12.48
```

```
Iteration 417/500: rewards 25.25 +/- 22.43
Iteration 418/500: rewards 23.38 +/- 11.95
Iteration 419/500: rewards 23.41 +/- 8.99
Iteration 420/500: rewards 22.12 +/- 8.99
Iteration 421/500: rewards 21.0 +/- 8.6
Iteration 422/500: rewards 22.22 +/- 11.09
Iteration 423/500: rewards 22.53 +/- 14.33
Iteration 424/500: rewards 19.53 +/- 8.44
Iteration 425/500: rewards 22.38 +/- 12.62
Iteration 426/500: rewards 22.12 +/- 12.61
Iteration 427/500: rewards 21.19 +/- 10.73
Iteration 428/500: rewards 23.78 +/- 10.51
Iteration 429/500: rewards 23.75 +/- 12.16
Iteration 430/500: rewards 21.28 +/- 8.73
Iteration 431/500: rewards 19.88 +/- 8.14
Iteration 432/500: rewards 25.34 +/- 12.72
Iteration 433/500: rewards 22.31 +/- 11.78
Iteration 434/500: rewards 23.16 +/- 9.83
Iteration 435/500: rewards 27.28 +/- 20.49
Iteration 436/500: rewards 22.56 +/- 10.17
Iteration 437/500: rewards 26.16 +/- 12.19
Iteration 438/500: rewards 21.25 +/- 8.73
Iteration 439/500: rewards 20.59 +/- 7.12
Iteration 440/500: rewards 24.38 +/- 15.91
Iteration 441/500: rewards 22.75 +/- 11.47
Iteration 442/500: rewards 22.69 +/- 10.41
Iteration 443/500: rewards 19.25 +/- 8.54
Iteration 444/500: rewards 20.69 +/- 8.17
Iteration 445/500: rewards 21.38 +/- 12.32
Iteration 446/500: rewards 21.22 +/- 13.09
Iteration 447/500: rewards 25.06 +/- 13.52
Iteration 448/500: rewards 23.75 +/- 14.08
Iteration 449/500: rewards 23.34 +/- 10.29
Iteration 450/500: rewards 25.16 +/- 16.06
Iteration 451/500: rewards 24.22 +/- 13.16
Iteration 452/500: rewards 33.03 +/- 28.0
Iteration 453/500: rewards 22.56 +/- 9.82
Iteration 454/500: rewards 22.47 +/- 10.4
Iteration 455/500: rewards 22.84 +/- 13.49
Iteration 456/500: rewards 28.41 +/- 16.76
Iteration 457/500: rewards 26.66 +/- 17.5
Iteration 458/500: rewards 29.81 +/- 22.32
Iteration 459/500: rewards 21.12 +/- 8.67
Iteration 460/500: rewards 24.91 +/- 12.47
Iteration 461/500: rewards 27.09 +/- 17.73
Iteration 462/500: rewards 24.06 +/- 12.62
Iteration 463/500: rewards 25.47 +/- 15.15
Iteration 464/500: rewards 23.66 +/- 16.54
```

```
Iteration 465/500: rewards 21.53 +/- 10.77
Iteration 466/500: rewards 21.78 +/- 9.15
Iteration 467/500: rewards 23.72 +/- 12.12
Iteration 468/500: rewards 22.62 +/- 12.5
Iteration 469/500: rewards 20.47 +/- 10.37
Iteration 470/500: rewards 24.03 +/- 12.37
Iteration 471/500: rewards 24.69 +/- 14.07
Iteration 472/500: rewards 22.91 +/- 12.12
Iteration 473/500: rewards 25.31 +/- 11.83
Iteration 474/500: rewards 24.94 +/- 10.56
Iteration 475/500: rewards 20.28 +/- 9.1
Iteration 476/500: rewards 25.5 +/- 11.63
Iteration 477/500: rewards 21.12 +/- 8.84
Iteration 478/500: rewards 21.97 +/- 9.52
Iteration 479/500: rewards 24.56 +/- 14.16
Iteration 480/500: rewards 24.62 +/- 16.02
Iteration 481/500: rewards 29.16 +/- 17.62
Iteration 482/500: rewards 25.28 +/- 15.01
Iteration 483/500: rewards 26.12 +/- 13.86
Iteration 484/500: rewards 23.53 +/- 10.26
Iteration 485/500: rewards 21.22 +/- 9.55
Iteration 486/500: rewards 25.97 +/- 17.06
Iteration 487/500: rewards 19.81 +/- 7.74
Iteration 488/500: rewards 27.38 +/- 15.39
Iteration 489/500: rewards 22.62 +/- 9.22
Iteration 490/500: rewards 24.34 +/- 11.3
Iteration 491/500: rewards 26.84 +/- 18.7
Iteration 492/500: rewards 22.06 +/- 8.51
Iteration 493/500: rewards 24.94 +/- 14.13
Iteration 494/500: rewards 21.38 +/- 7.62
Iteration 495/500: rewards 26.19 +/- 16.82
Iteration 496/500: rewards 26.59 +/- 14.62
Iteration 497/500: rewards 26.88 +/- 13.93
Iteration 498/500: rewards 23.81 +/- 12.8
Iteration 499/500: rewards 21.12 +/- 8.88
Iteration 500/500: rewards 24.12 +/- 15.81
The average reward is 24.315
the device is: cpu

------------------------------------------------
The gamma chosen is: 0.95
The value lr chosen is 0.001
The policy lr chosen is 0.0001
Iteration 1/500: rewards 19.0 +/- 7.26
Iteration 2/500: rewards 18.59 +/- 9.64
Iteration 3/500: rewards 17.69 +/- 8.52
Iteration 4/500: rewards 20.84 +/- 11.63
Iteration 5/500: rewards 17.03 +/- 8.93
Iteration 6/500: rewards 19.59 +/- 9.46
```

```
Iteration 7/500: rewards 17.94 +/- 9.8
Iteration 8/500: rewards 18.69 +/- 7.56
Iteration 9/500: rewards 19.84 +/- 7.51
Iteration 10/500: rewards 20.5 +/- 8.17
Iteration 11/500: rewards 20.5 +/- 8.84
Iteration 12/500: rewards 24.31 +/- 16.31
Iteration 13/500: rewards 19.78 +/- 10.8
Iteration 14/500: rewards 18.78 +/- 9.8
Iteration 15/500: rewards 19.47 +/- 13.23
Iteration 16/500: rewards 22.0 +/- 11.87
Iteration 17/500: rewards 19.38 +/- 9.61
Iteration 18/500: rewards 19.41 +/- 10.2
Iteration 19/500: rewards 20.25 +/- 9.54
Iteration 20/500: rewards 18.62 +/- 7.92
Iteration 21/500: rewards 19.75 +/- 9.2
Iteration 22/500: rewards 19.28 +/- 9.38
Iteration 23/500: rewards 16.31 +/- 7.02
Iteration 24/500: rewards 17.44 +/- 6.22
Iteration 25/500: rewards 17.56 +/- 7.87
Iteration 26/500: rewards 18.81 +/- 10.58
Iteration 27/500: rewards 19.44 +/- 7.07
Iteration 28/500: rewards 18.75 +/- 7.17
Iteration 29/500: rewards 19.12 +/- 9.72
Iteration 30/500: rewards 18.75 +/- 8.22
Iteration 31/500: rewards 16.66 +/- 7.18
Iteration 32/500: rewards 20.12 +/- 8.82
Iteration 33/500: rewards 18.94 +/- 12.44
Iteration 34/500: rewards 18.25 +/- 11.2
Iteration 35/500: rewards 22.28 +/- 16.03
Iteration 36/500: rewards 20.78 +/- 9.67
Iteration 37/500: rewards 17.62 +/- 7.26
Iteration 38/500: rewards 20.97 +/- 10.98
Iteration 39/500: rewards 14.94 +/- 5.88
Iteration 40/500: rewards 18.38 +/- 7.74
Iteration 41/500: rewards 17.84 +/- 8.3
Iteration 42/500: rewards 20.12 +/- 11.33
Iteration 43/500: rewards 16.88 +/- 5.67
Iteration 44/500: rewards 21.53 +/- 13.13
Iteration 45/500: rewards 22.25 +/- 11.29
Iteration 46/500: rewards 19.75 +/- 10.15
Iteration 47/500: rewards 21.91 +/- 11.95
Iteration 48/500: rewards 19.53 +/- 10.65
Iteration 49/500: rewards 19.81 +/- 8.94
Iteration 50/500: rewards 17.69 +/- 6.24
Iteration 51/500: rewards 23.47 +/- 16.83
Iteration 52/500: rewards 21.44 +/- 11.77
Iteration 53/500: rewards 18.0 +/- 7.04
Iteration 54/500: rewards 17.56 +/- 8.52
```

```
Iteration 55/500: rewards 18.28 +/- 7.84
Iteration 56/500: rewards 20.06 +/- 11.59
Iteration 57/500: rewards 22.62 +/- 11.64
Iteration 58/500: rewards 19.31 +/- 11.13
Iteration 59/500: rewards 16.31 +/- 6.82
Iteration 60/500: rewards 18.66 +/- 9.0
Iteration 61/500: rewards 18.0 +/- 8.3
Iteration 62/500: rewards 18.59 +/- 6.51
Iteration 63/500: rewards 21.19 +/- 9.99
Iteration 64/500: rewards 22.03 +/- 10.86
Iteration 65/500: rewards 18.81 +/- 8.54
Iteration 66/500: rewards 20.16 +/- 8.99
Iteration 67/500: rewards 23.97 +/- 15.09
Iteration 68/500: rewards 19.62 +/- 10.45
Iteration 69/500: rewards 18.53 +/- 6.38
Iteration 70/500: rewards 17.5 +/- 7.47
Iteration 71/500: rewards 22.53 +/- 12.92
Iteration 72/500: rewards 18.09 +/- 6.41
Iteration 73/500: rewards 17.12 +/- 6.76
Iteration 74/500: rewards 20.38 +/- 11.39
Iteration 75/500: rewards 23.16 +/- 12.71
Iteration 76/500: rewards 19.28 +/- 10.44
Iteration 77/500: rewards 20.28 +/- 9.92
Iteration 78/500: rewards 19.03 +/- 10.12
Iteration 79/500: rewards 21.91 +/- 12.63
Iteration 80/500: rewards 17.84 +/- 7.3
Iteration 81/500: rewards 17.0 +/- 7.31
Iteration 82/500: rewards 18.03 +/- 7.93
Iteration 83/500: rewards 21.56 +/- 11.69
Iteration 84/500: rewards 18.31 +/- 9.2
Iteration 85/500: rewards 16.53 +/- 7.91
Iteration 86/500: rewards 20.0 +/- 11.51
Iteration 87/500: rewards 20.38 +/- 7.71
Iteration 88/500: rewards 17.31 +/- 8.81
Iteration 89/500: rewards 19.97 +/- 11.14
Iteration 90/500: rewards 20.62 +/- 8.35
Iteration 91/500: rewards 20.12 +/- 9.22
Iteration 92/500: rewards 19.03 +/- 7.75
Iteration 93/500: rewards 20.94 +/- 12.25
Iteration 94/500: rewards 21.16 +/- 12.41
Iteration 95/500: rewards 20.94 +/- 9.42
Iteration 96/500: rewards 19.69 +/- 9.64
Iteration 97/500: rewards 18.81 +/- 8.46
Iteration 98/500: rewards 18.03 +/- 7.96
Iteration 99/500: rewards 20.09 +/- 10.06
Iteration 100/500: rewards 19.75 +/- 8.63
Iteration 101/500: rewards 18.03 +/- 6.88
Iteration 102/500: rewards 21.03 +/- 10.04
```

```
Iteration 103/500: rewards 21.19 +/- 11.95
Iteration 104/500: rewards 18.88 +/- 9.41
Iteration 105/500: rewards 21.69 +/- 9.65
Iteration 106/500: rewards 21.06 +/- 13.46
Iteration 107/500: rewards 19.78 +/- 9.12
Iteration 108/500: rewards 19.03 +/- 7.26
Iteration 109/500: rewards 17.34 +/- 7.04
Iteration 110/500: rewards 21.22 +/- 10.6
Iteration 111/500: rewards 19.66 +/- 11.35
Iteration 112/500: rewards 16.41 +/- 6.81
Iteration 113/500: rewards 19.56 +/- 9.14
Iteration 114/500: rewards 21.09 +/- 10.83
Iteration 115/500: rewards 18.12 +/- 8.64
Iteration 116/500: rewards 18.91 +/- 10.32
Iteration 117/500: rewards 19.56 +/- 11.76
Iteration 118/500: rewards 18.75 +/- 9.66
Iteration 119/500: rewards 18.25 +/- 7.99
Iteration 120/500: rewards 19.16 +/- 8.24
Iteration 121/500: rewards 14.81 +/- 5.44
Iteration 122/500: rewards 19.44 +/- 10.13
Iteration 123/500: rewards 18.47 +/- 8.99
Iteration 124/500: rewards 19.09 +/- 8.31
Iteration 125/500: rewards 20.66 +/- 12.47
Iteration 126/500: rewards 21.25 +/- 9.99
Iteration 127/500: rewards 21.31 +/- 12.16
Iteration 128/500: rewards 22.12 +/- 9.61
Iteration 129/500: rewards 17.28 +/- 5.64
Iteration 130/500: rewards 17.31 +/- 7.98
Iteration 131/500: rewards 23.59 +/- 14.63
Iteration 132/500: rewards 19.56 +/- 9.16
Iteration 133/500: rewards 19.38 +/- 7.02
Iteration 134/500: rewards 17.78 +/- 7.24
Iteration 135/500: rewards 21.09 +/- 11.38
Iteration 136/500: rewards 18.41 +/- 7.17
Iteration 137/500: rewards 20.19 +/- 8.89
Iteration 138/500: rewards 19.53 +/- 7.68
Iteration 139/500: rewards 20.41 +/- 13.2
Iteration 140/500: rewards 20.69 +/- 13.45
Iteration 141/500: rewards 20.31 +/- 10.48
Iteration 142/500: rewards 18.78 +/- 7.19
Iteration 143/500: rewards 19.59 +/- 13.2
Iteration 144/500: rewards 20.12 +/- 10.24
Iteration 145/500: rewards 21.28 +/- 10.73
Iteration 146/500: rewards 20.44 +/- 9.33
Iteration 147/500: rewards 19.56 +/- 12.37
Iteration 148/500: rewards 20.75 +/- 9.37
Iteration 149/500: rewards 20.88 +/- 14.28
Iteration 150/500: rewards 18.34 +/- 8.6
```

```
Iteration 151/500: rewards 16.59 +/- 7.18
Iteration 152/500: rewards 21.0 +/- 10.41
Iteration 153/500: rewards 17.72 +/- 6.45
Iteration 154/500: rewards 22.84 +/- 14.62
Iteration 155/500: rewards 18.09 +/- 6.85
Iteration 156/500: rewards 18.06 +/- 7.39
Iteration 157/500: rewards 20.25 +/- 9.85
Iteration 158/500: rewards 20.56 +/- 9.34
Iteration 159/500: rewards 20.62 +/- 12.45
Iteration 160/500: rewards 19.47 +/- 8.44
Iteration 161/500: rewards 18.62 +/- 6.69
Iteration 162/500: rewards 21.06 +/- 11.46
Iteration 163/500: rewards 19.22 +/- 8.56
Iteration 164/500: rewards 18.03 +/- 7.45
Iteration 165/500: rewards 18.34 +/- 5.95
Iteration 166/500: rewards 22.06 +/- 11.41
Iteration 167/500: rewards 19.97 +/- 10.37
Iteration 168/500: rewards 18.47 +/- 8.51
Iteration 169/500: rewards 19.16 +/- 9.16
Iteration 170/500: rewards 20.28 +/- 9.91
Iteration 171/500: rewards 17.09 +/- 6.1
Iteration 172/500: rewards 16.94 +/- 7.47
Iteration 173/500: rewards 22.38 +/- 10.42
Iteration 174/500: rewards 18.56 +/- 8.08
Iteration 175/500: rewards 21.62 +/- 13.66
Iteration 176/500: rewards 18.22 +/- 9.09
Iteration 177/500: rewards 18.88 +/- 8.76
Iteration 178/500: rewards 18.56 +/- 6.32
Iteration 179/500: rewards 20.31 +/- 7.47
Iteration 180/500: rewards 16.47 +/- 6.95
Iteration 181/500: rewards 21.16 +/- 8.72
Iteration 182/500: rewards 17.12 +/- 7.31
Iteration 183/500: rewards 19.38 +/- 8.22
Iteration 184/500: rewards 18.19 +/- 8.33
Iteration 185/500: rewards 20.91 +/- 10.09
Iteration 186/500: rewards 18.84 +/- 9.19
Iteration 187/500: rewards 19.12 +/- 7.27
Iteration 188/500: rewards 18.03 +/- 7.22
Iteration 189/500: rewards 18.94 +/- 8.87
Iteration 190/500: rewards 19.75 +/- 11.89
Iteration 191/500: rewards 17.34 +/- 6.22
Iteration 192/500: rewards 22.53 +/- 12.16
Iteration 193/500: rewards 21.72 +/- 13.81
Iteration 194/500: rewards 21.53 +/- 9.99
Iteration 195/500: rewards 20.03 +/- 10.16
Iteration 196/500: rewards 17.56 +/- 6.76
Iteration 197/500: rewards 19.38 +/- 8.25
Iteration 198/500: rewards 20.47 +/- 9.85
```

```
Iteration 199/500: rewards 18.38 +/- 6.69
Iteration 200/500: rewards 20.19 +/- 9.55
Iteration 201/500: rewards 20.0 +/- 9.8
Iteration 202/500: rewards 16.66 +/- 7.09
Iteration 203/500: rewards 21.66 +/- 14.61
Iteration 204/500: rewards 23.31 +/- 10.69
Iteration 205/500: rewards 18.41 +/- 8.05
Iteration 206/500: rewards 18.94 +/- 9.59
Iteration 207/500: rewards 16.75 +/- 8.86
Iteration 208/500: rewards 16.62 +/- 5.37
Iteration 209/500: rewards 20.88 +/- 12.51
Iteration 210/500: rewards 19.94 +/- 8.29
Iteration 211/500: rewards 17.69 +/- 9.06
Iteration 212/500: rewards 22.0 +/- 11.28
Iteration 213/500: rewards 19.56 +/- 10.17
Iteration 214/500: rewards 17.62 +/- 8.7
Iteration 215/500: rewards 22.25 +/- 10.8
Iteration 216/500: rewards 19.72 +/- 10.29
Iteration 217/500: rewards 20.16 +/- 10.56
Iteration 218/500: rewards 18.25 +/- 10.42
Iteration 219/500: rewards 16.44 +/- 5.28
Iteration 220/500: rewards 18.62 +/- 9.73
Iteration 221/500: rewards 19.69 +/- 10.28
Iteration 222/500: rewards 19.34 +/- 9.15
Iteration 223/500: rewards 19.28 +/- 7.8
Iteration 224/500: rewards 20.62 +/- 9.88
Iteration 225/500: rewards 18.16 +/- 7.84
Iteration 226/500: rewards 19.78 +/- 7.99
Iteration 227/500: rewards 19.78 +/- 9.56
Iteration 228/500: rewards 19.12 +/- 6.11
Iteration 229/500: rewards 20.78 +/- 9.83
Iteration 230/500: rewards 16.47 +/- 7.48
Iteration 231/500: rewards 22.84 +/- 12.99
Iteration 232/500: rewards 20.75 +/- 9.34
Iteration 233/500: rewards 18.66 +/- 7.44
Iteration 234/500: rewards 22.25 +/- 13.01
Iteration 235/500: rewards 19.78 +/- 7.47
Iteration 236/500: rewards 20.06 +/- 9.41
Iteration 237/500: rewards 23.0 +/- 16.92
Iteration 238/500: rewards 21.84 +/- 10.73
Iteration 239/500: rewards 21.25 +/- 11.48
Iteration 240/500: rewards 22.0 +/- 12.17
Iteration 241/500: rewards 18.91 +/- 8.14
Iteration 242/500: rewards 20.94 +/- 10.48
Iteration 243/500: rewards 19.62 +/- 9.54
Iteration 244/500: rewards 18.59 +/- 6.76
Iteration 245/500: rewards 23.09 +/- 14.7
Iteration 246/500: rewards 20.25 +/- 9.71
```

```
Iteration 247/500: rewards 19.94 +/- 12.23
Iteration 248/500: rewards 19.72 +/- 8.51
Iteration 249/500: rewards 16.91 +/- 6.34
Iteration 250/500: rewards 22.91 +/- 15.55
Iteration 251/500: rewards 18.28 +/- 7.4
Iteration 252/500: rewards 20.94 +/- 9.23
Iteration 253/500: rewards 21.56 +/- 10.62
Iteration 254/500: rewards 21.94 +/- 10.64
Iteration 255/500: rewards 20.81 +/- 8.64
Iteration 256/500: rewards 23.84 +/- 16.14
Iteration 257/500: rewards 19.12 +/- 9.52
Iteration 258/500: rewards 18.94 +/- 9.25
Iteration 259/500: rewards 17.5 +/- 6.1
Iteration 260/500: rewards 19.03 +/- 8.96
Iteration 261/500: rewards 20.0 +/- 8.72
Iteration 262/500: rewards 21.12 +/- 7.49
Iteration 263/500: rewards 19.78 +/- 7.17
Iteration 264/500: rewards 19.97 +/- 8.7
Iteration 265/500: rewards 18.47 +/- 9.66
Iteration 266/500: rewards 18.62 +/- 8.38
Iteration 267/500: rewards 21.53 +/- 9.26
Iteration 268/500: rewards 18.56 +/- 9.7
Iteration 269/500: rewards 20.03 +/- 8.71
Iteration 270/500: rewards 21.5 +/- 12.35
Iteration 271/500: rewards 18.75 +/- 8.67
Iteration 272/500: rewards 18.91 +/- 7.74
Iteration 273/500: rewards 20.94 +/- 9.39
Iteration 274/500: rewards 18.62 +/- 6.02
Iteration 275/500: rewards 20.62 +/- 12.05
Iteration 276/500: rewards 23.62 +/- 12.39
Iteration 277/500: rewards 23.22 +/- 12.25
Iteration 278/500: rewards 21.75 +/- 12.68
Iteration 279/500: rewards 21.34 +/- 10.77
Iteration 280/500: rewards 19.12 +/- 8.07
Iteration 281/500: rewards 23.69 +/- 14.33
Iteration 282/500: rewards 22.62 +/- 13.24
Iteration 283/500: rewards 18.34 +/- 6.48
Iteration 284/500: rewards 18.53 +/- 7.56
Iteration 285/500: rewards 23.22 +/- 11.37
Iteration 286/500: rewards 20.91 +/- 12.18
Iteration 287/500: rewards 21.5 +/- 8.24
Iteration 288/500: rewards 22.56 +/- 13.82
Iteration 289/500: rewards 23.72 +/- 10.75
Iteration 290/500: rewards 23.53 +/- 12.7
Iteration 291/500: rewards 20.75 +/- 6.9
Iteration 292/500: rewards 21.81 +/- 9.83
Iteration 293/500: rewards 23.5 +/- 9.43
Iteration 294/500: rewards 22.75 +/- 9.03
```

```
Iteration 295/500: rewards 20.5 +/- 11.27
Iteration 296/500: rewards 22.56 +/- 9.16
Iteration 297/500: rewards 17.75 +/- 7.34
Iteration 298/500: rewards 17.34 +/- 7.66
Iteration 299/500: rewards 21.66 +/- 11.64
Iteration 300/500: rewards 21.81 +/- 11.01
Iteration 301/500: rewards 18.81 +/- 8.02
Iteration 302/500: rewards 19.34 +/- 13.43
Iteration 303/500: rewards 20.62 +/- 7.32
Iteration 304/500: rewards 22.12 +/- 18.1
Iteration 305/500: rewards 23.06 +/- 11.63
Iteration 306/500: rewards 24.66 +/- 12.33
Iteration 307/500: rewards 20.62 +/- 9.09
Iteration 308/500: rewards 20.75 +/- 10.02
Iteration 309/500: rewards 20.16 +/- 9.17
Iteration 310/500: rewards 19.75 +/- 10.56
Iteration 311/500: rewards 20.88 +/- 10.29
Iteration 312/500: rewards 20.44 +/- 12.15
Iteration 313/500: rewards 23.97 +/- 18.57
Iteration 314/500: rewards 25.25 +/- 18.51
Iteration 315/500: rewards 21.09 +/- 10.37
Iteration 316/500: rewards 21.22 +/- 9.03
Iteration 317/500: rewards 23.03 +/- 11.5
Iteration 318/500: rewards 19.44 +/- 8.38
Iteration 319/500: rewards 18.69 +/- 8.68
Iteration 320/500: rewards 23.66 +/- 13.37
Iteration 321/500: rewards 21.72 +/- 10.66
Iteration 322/500: rewards 23.72 +/- 16.31
Iteration 323/500: rewards 20.62 +/- 9.97
Iteration 324/500: rewards 20.16 +/- 8.57
Iteration 325/500: rewards 25.5 +/- 16.31
Iteration 326/500: rewards 23.81 +/- 10.96
Iteration 327/500: rewards 22.22 +/- 10.58
Iteration 328/500: rewards 21.09 +/- 9.89
Iteration 329/500: rewards 21.22 +/- 10.93
Iteration 330/500: rewards 18.97 +/- 9.04
Iteration 331/500: rewards 21.75 +/- 9.12
Iteration 332/500: rewards 23.12 +/- 10.08
Iteration 333/500: rewards 20.25 +/- 11.59
Iteration 334/500: rewards 22.03 +/- 12.16
Iteration 335/500: rewards 19.16 +/- 5.03
Iteration 336/500: rewards 20.66 +/- 9.28
Iteration 337/500: rewards 22.28 +/- 12.36
Iteration 338/500: rewards 23.97 +/- 14.57
Iteration 339/500: rewards 21.06 +/- 11.48
Iteration 340/500: rewards 20.28 +/- 12.35
Iteration 341/500: rewards 21.5 +/- 11.6
Iteration 342/500: rewards 28.16 +/- 18.29
```

```
Iteration 343/500: rewards 24.69 +/- 16.92
Iteration 344/500: rewards 20.78 +/- 11.25
Iteration 345/500: rewards 22.59 +/- 13.36
Iteration 346/500: rewards 23.16 +/- 13.56
Iteration 347/500: rewards 19.66 +/- 6.19
Iteration 348/500: rewards 20.22 +/- 8.23
Iteration 349/500: rewards 19.03 +/- 9.14
Iteration 350/500: rewards 21.47 +/- 9.12
Iteration 351/500: rewards 21.16 +/- 13.03
Iteration 352/500: rewards 26.19 +/- 15.34
Iteration 353/500: rewards 18.81 +/- 9.32
Iteration 354/500: rewards 25.75 +/- 13.96
Iteration 355/500: rewards 19.22 +/- 7.42
Iteration 356/500: rewards 20.66 +/- 9.11
Iteration 357/500: rewards 20.91 +/- 9.65
Iteration 358/500: rewards 26.38 +/- 22.53
Iteration 359/500: rewards 26.5 +/- 15.55
Iteration 360/500: rewards 21.25 +/- 11.64
Iteration 361/500: rewards 21.78 +/- 10.61
Iteration 362/500: rewards 22.78 +/- 9.89
Iteration 363/500: rewards 25.06 +/- 13.77
Iteration 364/500: rewards 24.44 +/- 13.44
Iteration 365/500: rewards 22.38 +/- 18.73
Iteration 366/500: rewards 21.22 +/- 8.28
Iteration 367/500: rewards 25.03 +/- 19.94
Iteration 368/500: rewards 21.09 +/- 11.05
Iteration 369/500: rewards 19.66 +/- 7.26
Iteration 370/500: rewards 22.03 +/- 9.53
Iteration 371/500: rewards 23.53 +/- 11.08
Iteration 372/500: rewards 24.53 +/- 10.81
Iteration 373/500: rewards 20.47 +/- 14.18
Iteration 374/500: rewards 22.78 +/- 13.16
Iteration 375/500: rewards 20.66 +/- 10.02
Iteration 376/500: rewards 20.25 +/- 9.43
Iteration 377/500: rewards 21.38 +/- 10.1
Iteration 378/500: rewards 22.0 +/- 11.66
Iteration 379/500: rewards 21.12 +/- 10.25
Iteration 380/500: rewards 18.06 +/- 6.83
Iteration 381/500: rewards 20.97 +/- 9.02
Iteration 382/500: rewards 21.62 +/- 14.47
Iteration 383/500: rewards 20.94 +/- 8.27
Iteration 384/500: rewards 21.56 +/- 10.64
Iteration 385/500: rewards 27.19 +/- 19.73
Iteration 386/500: rewards 24.91 +/- 14.47
Iteration 387/500: rewards 21.78 +/- 9.91
Iteration 388/500: rewards 22.53 +/- 10.15
Iteration 389/500: rewards 21.03 +/- 12.97
Iteration 390/500: rewards 20.5 +/- 11.68
```

```
Iteration 391/500: rewards 24.94 +/- 16.21
Iteration 392/500: rewards 24.22 +/- 11.82
Iteration 393/500: rewards 27.34 +/- 13.84
Iteration 394/500: rewards 25.44 +/- 13.11
Iteration 395/500: rewards 22.5 +/- 10.9
Iteration 396/500: rewards 23.31 +/- 14.09
Iteration 397/500: rewards 21.56 +/- 10.53
Iteration 398/500: rewards 23.03 +/- 16.81
Iteration 399/500: rewards 25.0 +/- 13.15
Iteration 400/500: rewards 21.06 +/- 12.6
Iteration 401/500: rewards 21.44 +/- 10.24
Iteration 402/500: rewards 18.22 +/- 9.24
Iteration 403/500: rewards 20.41 +/- 12.67
Iteration 404/500: rewards 24.84 +/- 13.62
Iteration 405/500: rewards 20.66 +/- 9.4
Iteration 406/500: rewards 19.72 +/- 9.16
Iteration 407/500: rewards 22.28 +/- 11.02
Iteration 408/500: rewards 21.94 +/- 14.54
Iteration 409/500: rewards 20.31 +/- 7.99
Iteration 410/500: rewards 20.16 +/- 6.82
Iteration 411/500: rewards 20.25 +/- 8.74
Iteration 412/500: rewards 27.06 +/- 15.53
Iteration 413/500: rewards 23.16 +/- 11.26
Iteration 414/500: rewards 21.31 +/- 11.38
Iteration 415/500: rewards 22.47 +/- 11.54
Iteration 416/500: rewards 27.84 +/- 18.51
Iteration 417/500: rewards 22.97 +/- 15.36
Iteration 418/500: rewards 21.34 +/- 11.35
Iteration 419/500: rewards 24.12 +/- 12.81
Iteration 420/500: rewards 23.12 +/- 10.58
Iteration 421/500: rewards 24.12 +/- 10.25
Iteration 422/500: rewards 20.31 +/- 9.01
Iteration 423/500: rewards 24.97 +/- 13.21
Iteration 424/500: rewards 23.31 +/- 12.3
Iteration 425/500: rewards 25.06 +/- 13.15
Iteration 426/500: rewards 21.44 +/- 11.8
Iteration 427/500: rewards 19.72 +/- 7.86
Iteration 428/500: rewards 29.41 +/- 19.55
Iteration 429/500: rewards 23.69 +/- 11.02
Iteration 430/500: rewards 22.56 +/- 14.11
Iteration 431/500: rewards 22.06 +/- 15.95
Iteration 432/500: rewards 24.78 +/- 13.27
Iteration 433/500: rewards 24.0 +/- 19.26
Iteration 434/500: rewards 20.12 +/- 9.43
Iteration 435/500: rewards 22.03 +/- 10.8
Iteration 436/500: rewards 21.72 +/- 9.27
Iteration 437/500: rewards 25.16 +/- 14.58
Iteration 438/500: rewards 29.06 +/- 20.11
```

```
Iteration 439/500: rewards 22.66 +/- 16.72
Iteration 440/500: rewards 21.56 +/- 9.11
Iteration 441/500: rewards 25.31 +/- 17.49
Iteration 442/500: rewards 25.22 +/- 13.21
Iteration 443/500: rewards 21.53 +/- 11.51
Iteration 444/500: rewards 22.59 +/- 11.46
Iteration 445/500: rewards 24.69 +/- 13.47
Iteration 446/500: rewards 21.91 +/- 13.41
Iteration 447/500: rewards 21.81 +/- 10.02
Iteration 448/500: rewards 19.62 +/- 6.79
Iteration 449/500: rewards 23.28 +/- 14.57
Iteration 450/500: rewards 26.22 +/- 13.6
Iteration 451/500: rewards 23.22 +/- 18.58
Iteration 452/500: rewards 20.38 +/- 11.61
Iteration 453/500: rewards 20.41 +/- 9.68
Iteration 454/500: rewards 25.59 +/- 15.12
Iteration 455/500: rewards 21.62 +/- 8.57
Iteration 456/500: rewards 24.69 +/- 15.19
Iteration 457/500: rewards 21.22 +/- 9.14
Iteration 458/500: rewards 25.56 +/- 15.82
Iteration 459/500: rewards 24.5 +/- 13.65
Iteration 460/500: rewards 24.41 +/- 10.87
Iteration 461/500: rewards 22.94 +/- 14.68
Iteration 462/500: rewards 21.81 +/- 11.29
Iteration 463/500: rewards 25.41 +/- 14.3
Iteration 464/500: rewards 23.44 +/- 14.46
Iteration 465/500: rewards 20.88 +/- 9.42
Iteration 466/500: rewards 22.84 +/- 9.45
Iteration 467/500: rewards 24.81 +/- 13.54
Iteration 468/500: rewards 23.5 +/- 9.13
Iteration 469/500: rewards 25.91 +/- 13.77
Iteration 470/500: rewards 27.19 +/- 12.67
Iteration 471/500: rewards 21.0 +/- 10.8
Iteration 472/500: rewards 22.69 +/- 10.73
Iteration 473/500: rewards 22.81 +/- 12.37
Iteration 474/500: rewards 23.09 +/- 10.62
Iteration 475/500: rewards 21.06 +/- 9.6
Iteration 476/500: rewards 24.62 +/- 12.94
Iteration 477/500: rewards 22.34 +/- 7.97
Iteration 478/500: rewards 22.75 +/- 13.97
Iteration 479/500: rewards 26.31 +/- 13.71
Iteration 480/500: rewards 27.56 +/- 16.03
Iteration 481/500: rewards 26.5 +/- 18.65
Iteration 482/500: rewards 26.25 +/- 13.3
Iteration 483/500: rewards 24.81 +/- 13.91
Iteration 484/500: rewards 20.31 +/- 9.23
Iteration 485/500: rewards 26.72 +/- 13.99
Iteration 486/500: rewards 28.47 +/- 18.5
```

```
Iteration 487/500: rewards 28.19 +/- 17.75
Iteration 488/500: rewards 29.47 +/- 16.88
Iteration 489/500: rewards 32.78 +/- 15.64
Iteration 490/500: rewards 21.28 +/- 8.65
Iteration 491/500: rewards 24.31 +/- 11.25
Iteration 492/500: rewards 24.16 +/- 11.05
Iteration 493/500: rewards 24.88 +/- 13.01
Iteration 494/500: rewards 24.22 +/- 13.03
Iteration 495/500: rewards 22.47 +/- 10.12
Iteration 496/500: rewards 19.69 +/- 10.3
Iteration 497/500: rewards 27.12 +/- 16.36
Iteration 498/500: rewards 26.34 +/- 22.98
Iteration 499/500: rewards 21.81 +/- 10.94
Iteration 500/500: rewards 22.06 +/- 11.05
The average reward is 24.128125
the device is: cpu
--------------------------------------------------
The gamma chosen is: 0.95
The value lr chosen is 0.0001
The policy lr chosen is 0.0001
Iteration 1/500: rewards 19.0 +/- 7.26
Iteration 2/500: rewards 18.59 +/- 9.64
Iteration 3/500: rewards 17.69 +/- 8.52
Iteration 4/500: rewards 20.84 +/- 11.63
Iteration 5/500: rewards 17.03 +/- 8.93
Iteration 6/500: rewards 19.59 +/- 9.46
Iteration 7/500: rewards 17.94 +/- 9.8
Iteration 8/500: rewards 18.69 +/- 7.56
Iteration 9/500: rewards 19.84 +/- 7.51
Iteration 10/500: rewards 20.5 +/- 8.17
Iteration 11/500: rewards 20.5 +/- 8.84
Iteration 12/500: rewards 24.31 +/- 16.31
Iteration 13/500: rewards 19.78 +/- 10.8
Iteration 14/500: rewards 18.78 +/- 9.8
Iteration 15/500: rewards 19.47 +/- 13.23
Iteration 16/500: rewards 22.0 +/- 11.87
Iteration 17/500: rewards 19.38 +/- 9.61
Iteration 18/500: rewards 19.41 +/- 10.2
Iteration 19/500: rewards 20.25 +/- 9.54
Iteration 20/500: rewards 18.62 +/- 7.92
Iteration 21/500: rewards 19.75 +/- 9.2
Iteration 22/500: rewards 19.28 +/- 9.38
Iteration 23/500: rewards 16.31 +/- 7.02
Iteration 24/500: rewards 17.44 +/- 6.22
Iteration 25/500: rewards 17.56 +/- 7.87
Iteration 26/500: rewards 18.81 +/- 10.58
Iteration 27/500: rewards 19.44 +/- 7.07
Iteration 28/500: rewards 18.75 +/- 7.17
```

```
Iteration 29/500: rewards 19.12 +/- 9.72
Iteration 30/500: rewards 18.75 +/- 8.04
Iteration 31/500: rewards 16.66 +/- 7.18
Iteration 32/500: rewards 20.12 +/- 8.82
Iteration 33/500: rewards 20.53 +/- 12.22
Iteration 34/500: rewards 20.56 +/- 9.28
Iteration 35/500: rewards 21.38 +/- 10.22
Iteration 36/500: rewards 19.28 +/- 11.93
Iteration 37/500: rewards 20.66 +/- 10.5
Iteration 38/500: rewards 21.78 +/- 11.44
Iteration 39/500: rewards 17.06 +/- 7.83
Iteration 40/500: rewards 20.81 +/- 12.4
Iteration 41/500: rewards 17.66 +/- 6.53
Iteration 42/500: rewards 20.25 +/- 9.79
Iteration 43/500: rewards 22.09 +/- 11.15
Iteration 44/500: rewards 17.78 +/- 6.49
Iteration 45/500: rewards 23.28 +/- 13.27
Iteration 46/500: rewards 18.72 +/- 6.47
Iteration 47/500: rewards 19.88 +/- 9.28
Iteration 48/500: rewards 20.16 +/- 9.31
Iteration 49/500: rewards 18.69 +/- 8.41
Iteration 50/500: rewards 20.78 +/- 8.59
Iteration 51/500: rewards 18.06 +/- 7.42
Iteration 52/500: rewards 18.06 +/- 9.06
Iteration 53/500: rewards 17.5 +/- 6.78
Iteration 54/500: rewards 16.25 +/- 4.72
Iteration 55/500: rewards 19.53 +/- 8.74
Iteration 56/500: rewards 20.69 +/- 10.18
Iteration 57/500: rewards 21.5 +/- 11.61
Iteration 58/500: rewards 19.84 +/- 8.64
Iteration 59/500: rewards 17.59 +/- 7.7
Iteration 60/500: rewards 17.22 +/- 8.44
Iteration 61/500: rewards 22.03 +/- 12.13
Iteration 62/500: rewards 20.84 +/- 8.16
Iteration 63/500: rewards 20.28 +/- 13.25
Iteration 64/500: rewards 21.0 +/- 10.61
Iteration 65/500: rewards 17.78 +/- 6.45
Iteration 66/500: rewards 17.12 +/- 7.02
Iteration 67/500: rewards 17.16 +/- 6.36
Iteration 68/500: rewards 16.22 +/- 6.22
Iteration 69/500: rewards 18.0 +/- 10.11
Iteration 70/500: rewards 19.41 +/- 10.13
Iteration 71/500: rewards 16.72 +/- 6.3
Iteration 72/500: rewards 22.97 +/- 14.63
Iteration 73/500: rewards 19.34 +/- 10.74
Iteration 74/500: rewards 19.81 +/- 9.58
Iteration 75/500: rewards 17.09 +/- 6.34
Iteration 76/500: rewards 18.94 +/- 7.44
```

```
Iteration 77/500: rewards 20.69 +/- 10.11
Iteration 78/500: rewards 19.16 +/- 7.87
Iteration 79/500: rewards 21.5 +/- 10.78
Iteration 80/500: rewards 22.75 +/- 12.31
Iteration 81/500: rewards 17.03 +/- 6.47
Iteration 82/500: rewards 18.59 +/- 7.31
Iteration 83/500: rewards 17.66 +/- 8.48
Iteration 84/500: rewards 21.06 +/- 14.14
Iteration 85/500: rewards 15.34 +/- 6.04
Iteration 86/500: rewards 19.44 +/- 8.34
Iteration 87/500: rewards 21.34 +/- 10.82
Iteration 88/500: rewards 17.94 +/- 7.67
Iteration 89/500: rewards 20.22 +/- 12.11
Iteration 90/500: rewards 20.41 +/- 9.84
Iteration 91/500: rewards 18.94 +/- 7.79
Iteration 92/500: rewards 21.53 +/- 8.79
Iteration 93/500: rewards 19.0 +/- 9.12
Iteration 94/500: rewards 20.22 +/- 10.88
Iteration 95/500: rewards 19.41 +/- 12.41
Iteration 96/500: rewards 19.66 +/- 9.6
Iteration 97/500: rewards 18.69 +/- 7.09
Iteration 98/500: rewards 18.88 +/- 9.27
Iteration 99/500: rewards 21.34 +/- 10.16
Iteration 100/500: rewards 19.31 +/- 9.82
Iteration 101/500: rewards 21.53 +/- 11.33
Iteration 102/500: rewards 19.66 +/- 9.94
Iteration 103/500: rewards 17.94 +/- 7.93
Iteration 104/500: rewards 22.0 +/- 12.8
Iteration 105/500: rewards 22.38 +/- 13.67
Iteration 106/500: rewards 18.5 +/- 8.81
Iteration 107/500: rewards 23.28 +/- 14.29
Iteration 108/500: rewards 18.19 +/- 7.8
Iteration 109/500: rewards 17.16 +/- 7.56
Iteration 110/500: rewards 17.25 +/- 6.58
Iteration 111/500: rewards 21.84 +/- 11.58
Iteration 112/500: rewards 17.94 +/- 8.02
Iteration 113/500: rewards 19.22 +/- 9.32
Iteration 114/500: rewards 20.31 +/- 11.16
Iteration 115/500: rewards 18.72 +/- 8.46
Iteration 116/500: rewards 19.0 +/- 8.56
Iteration 117/500: rewards 22.59 +/- 12.55
Iteration 118/500: rewards 18.28 +/- 6.91
Iteration 119/500: rewards 18.78 +/- 8.93
Iteration 120/500: rewards 19.56 +/- 9.55
Iteration 121/500: rewards 15.97 +/- 6.04
Iteration 122/500: rewards 21.03 +/- 11.35
Iteration 123/500: rewards 20.5 +/- 10.66
Iteration 124/500: rewards 20.84 +/- 8.3
```

```
Iteration 125/500: rewards 22.0 +/- 12.15
Iteration 126/500: rewards 19.91 +/- 10.37
Iteration 127/500: rewards 20.22 +/- 12.22
Iteration 128/500: rewards 20.31 +/- 9.93
Iteration 129/500: rewards 17.12 +/- 7.41
Iteration 130/500: rewards 20.0 +/- 10.6
Iteration 131/500: rewards 21.12 +/- 9.18
Iteration 132/500: rewards 23.03 +/- 13.03
Iteration 133/500: rewards 17.25 +/- 6.34
Iteration 134/500: rewards 17.38 +/- 7.36
Iteration 135/500: rewards 21.19 +/- 9.91
Iteration 136/500: rewards 18.72 +/- 6.02
Iteration 137/500: rewards 21.75 +/- 12.93
Iteration 138/500: rewards 21.34 +/- 8.74
Iteration 139/500: rewards 18.44 +/- 9.93
Iteration 140/500: rewards 19.84 +/- 11.07
Iteration 141/500: rewards 20.06 +/- 9.11
Iteration 142/500: rewards 18.97 +/- 9.04
Iteration 143/500: rewards 18.25 +/- 6.95
Iteration 144/500: rewards 20.0 +/- 10.56
Iteration 145/500: rewards 20.34 +/- 12.04
Iteration 146/500: rewards 21.03 +/- 10.52
Iteration 147/500: rewards 18.56 +/- 6.75
Iteration 148/500: rewards 19.19 +/- 8.1
Iteration 149/500: rewards 18.19 +/- 7.7
Iteration 150/500: rewards 20.97 +/- 10.92
Iteration 151/500: rewards 16.47 +/- 6.55
Iteration 152/500: rewards 18.44 +/- 7.6
Iteration 153/500: rewards 18.62 +/- 7.36
Iteration 154/500: rewards 21.28 +/- 13.21
Iteration 155/500: rewards 19.53 +/- 10.58
Iteration 156/500: rewards 17.28 +/- 5.13
Iteration 157/500: rewards 20.03 +/- 7.61
Iteration 158/500: rewards 19.59 +/- 11.74
Iteration 159/500: rewards 21.28 +/- 10.23
Iteration 160/500: rewards 21.5 +/- 13.63
Iteration 161/500: rewards 17.94 +/- 7.71
Iteration 162/500: rewards 24.44 +/- 14.53
Iteration 163/500: rewards 18.69 +/- 5.72
Iteration 164/500: rewards 20.0 +/- 12.91
Iteration 165/500: rewards 20.75 +/- 7.62
Iteration 166/500: rewards 19.06 +/- 8.92
Iteration 167/500: rewards 20.16 +/- 9.77
Iteration 168/500: rewards 20.5 +/- 10.15
Iteration 169/500: rewards 19.56 +/- 9.11
Iteration 170/500: rewards 17.91 +/- 4.98
Iteration 171/500: rewards 18.72 +/- 7.69
Iteration 172/500: rewards 19.69 +/- 11.14
```

```
Iteration 173/500: rewards 20.44 +/- 6.8
Iteration 174/500: rewards 18.97 +/- 9.36
Iteration 175/500: rewards 19.69 +/- 9.07
Iteration 176/500: rewards 20.59 +/- 10.76
Iteration 177/500: rewards 22.19 +/- 12.53
Iteration 178/500: rewards 22.62 +/- 12.1
Iteration 179/500: rewards 20.0 +/- 10.05
Iteration 180/500: rewards 20.94 +/- 11.43
Iteration 181/500: rewards 20.22 +/- 12.82
Iteration 182/500: rewards 20.59 +/- 13.16
Iteration 183/500: rewards 18.41 +/- 6.5
Iteration 184/500: rewards 19.69 +/- 9.0
Iteration 185/500: rewards 19.12 +/- 8.03
Iteration 186/500: rewards 16.78 +/- 5.88
Iteration 187/500: rewards 18.03 +/- 7.58
Iteration 188/500: rewards 23.5 +/- 15.2
Iteration 189/500: rewards 20.62 +/- 10.91
Iteration 190/500: rewards 18.34 +/- 7.88
Iteration 191/500: rewards 21.69 +/- 11.33
Iteration 192/500: rewards 20.25 +/- 8.01
Iteration 193/500: rewards 20.47 +/- 10.09
Iteration 194/500: rewards 19.22 +/- 6.09
Iteration 195/500: rewards 20.41 +/- 10.64
Iteration 196/500: rewards 18.94 +/- 9.41
Iteration 197/500: rewards 19.09 +/- 8.57
Iteration 198/500: rewards 20.31 +/- 7.21
Iteration 199/500: rewards 22.56 +/- 12.17
Iteration 200/500: rewards 18.81 +/- 10.1
Iteration 201/500: rewards 18.94 +/- 7.41
Iteration 202/500: rewards 18.0 +/- 9.45
Iteration 203/500: rewards 21.0 +/- 9.16
Iteration 204/500: rewards 21.0 +/- 11.98
Iteration 205/500: rewards 22.38 +/- 10.6
Iteration 206/500: rewards 19.5 +/- 8.58
Iteration 207/500: rewards 20.53 +/- 13.21
Iteration 208/500: rewards 24.53 +/- 13.83
Iteration 209/500: rewards 19.06 +/- 9.89
Iteration 210/500: rewards 20.47 +/- 9.61
Iteration 211/500: rewards 20.12 +/- 7.51
Iteration 212/500: rewards 19.75 +/- 11.72
Iteration 213/500: rewards 20.44 +/- 9.58
Iteration 214/500: rewards 17.62 +/- 8.55
Iteration 215/500: rewards 25.16 +/- 16.34
Iteration 216/500: rewards 19.19 +/- 12.59
Iteration 217/500: rewards 15.53 +/- 6.44
Iteration 218/500: rewards 19.41 +/- 10.11
Iteration 219/500: rewards 19.97 +/- 8.21
Iteration 220/500: rewards 21.16 +/- 8.85
```

```
Iteration 221/500: rewards 19.31 +/- 8.83
Iteration 222/500: rewards 21.31 +/- 11.08
Iteration 223/500: rewards 19.41 +/- 8.91
Iteration 224/500: rewards 20.03 +/- 10.24
Iteration 225/500: rewards 18.0 +/- 8.26
Iteration 226/500: rewards 19.56 +/- 7.71
Iteration 227/500: rewards 17.28 +/- 7.2
Iteration 228/500: rewards 17.94 +/- 6.42
Iteration 229/500: rewards 21.44 +/- 11.37
Iteration 230/500: rewards 18.53 +/- 8.58
Iteration 231/500: rewards 20.88 +/- 9.88
Iteration 232/500: rewards 21.97 +/- 9.63
Iteration 233/500: rewards 18.81 +/- 9.27
Iteration 234/500: rewards 17.09 +/- 6.76
Iteration 235/500: rewards 18.38 +/- 11.03
Iteration 236/500: rewards 19.25 +/- 6.81
Iteration 237/500: rewards 20.0 +/- 8.67
Iteration 238/500: rewards 18.38 +/- 8.26
Iteration 239/500: rewards 21.5 +/- 9.95
Iteration 240/500: rewards 18.81 +/- 9.28
Iteration 241/500: rewards 17.69 +/- 8.16
Iteration 242/500: rewards 21.56 +/- 15.34
Iteration 243/500: rewards 19.72 +/- 11.87
Iteration 244/500: rewards 18.66 +/- 8.47
Iteration 245/500: rewards 18.16 +/- 6.65
Iteration 246/500: rewards 19.22 +/- 9.35
Iteration 247/500: rewards 21.22 +/- 8.23
Iteration 248/500: rewards 21.12 +/- 11.88
Iteration 249/500: rewards 20.88 +/- 9.17
Iteration 250/500: rewards 20.25 +/- 9.29
Iteration 251/500: rewards 20.47 +/- 9.71
Iteration 252/500: rewards 17.47 +/- 7.65
Iteration 253/500: rewards 19.31 +/- 9.11
Iteration 254/500: rewards 21.84 +/- 8.57
Iteration 255/500: rewards 20.34 +/- 8.83
Iteration 256/500: rewards 17.72 +/- 5.73
Iteration 257/500: rewards 16.03 +/- 4.97
Iteration 258/500: rewards 22.75 +/- 11.23
Iteration 259/500: rewards 22.69 +/- 13.44
Iteration 260/500: rewards 19.78 +/- 11.17
Iteration 261/500: rewards 20.28 +/- 10.33
Iteration 262/500: rewards 19.5 +/- 9.93
Iteration 263/500: rewards 18.47 +/- 7.08
Iteration 264/500: rewards 18.09 +/- 6.35
Iteration 265/500: rewards 20.41 +/- 11.45
Iteration 266/500: rewards 16.81 +/- 6.35
Iteration 267/500: rewards 20.34 +/- 11.51
Iteration 268/500: rewards 18.88 +/- 7.55
```

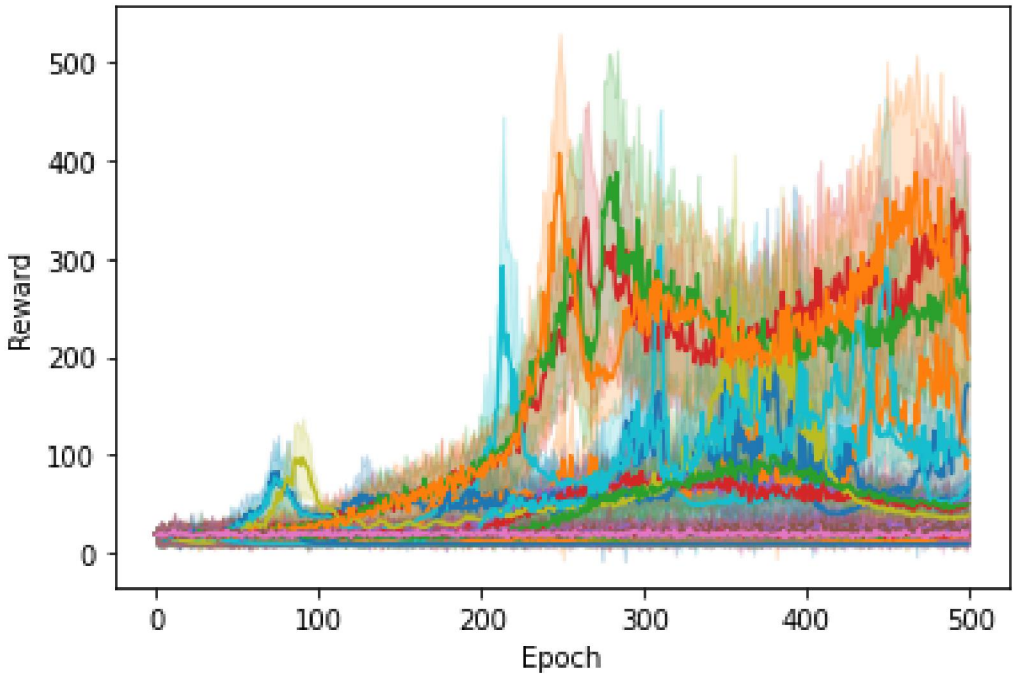
```
Iteration 269/500: rewards 18.78 +/- 10.49
Iteration 270/500: rewards 20.0 +/- 9.99
Iteration 271/500: rewards 21.75 +/- 12.94
Iteration 272/500: rewards 19.19 +/- 8.36
Iteration 273/500: rewards 17.47 +/- 5.74
Iteration 274/500: rewards 18.5 +/- 7.88
Iteration 275/500: rewards 22.41 +/- 17.23
Iteration 276/500: rewards 20.5 +/- 11.29
Iteration 277/500: rewards 20.0 +/- 10.47
Iteration 278/500: rewards 21.69 +/- 9.72
Iteration 279/500: rewards 16.81 +/- 5.7
Iteration 280/500: rewards 18.06 +/- 6.74
Iteration 281/500: rewards 21.59 +/- 12.1
Iteration 282/500: rewards 18.94 +/- 7.53
Iteration 283/500: rewards 18.5 +/- 6.93
Iteration 284/500: rewards 19.44 +/- 6.45
Iteration 285/500: rewards 16.34 +/- 6.52
Iteration 286/500: rewards 21.06 +/- 11.5
Iteration 287/500: rewards 20.34 +/- 8.72
Iteration 288/500: rewards 18.0 +/- 8.18
Iteration 289/500: rewards 18.97 +/- 9.67
Iteration 290/500: rewards 21.03 +/- 9.12
Iteration 291/500: rewards 22.06 +/- 9.38
Iteration 292/500: rewards 19.12 +/- 10.75
Iteration 293/500: rewards 19.19 +/- 7.82
Iteration 294/500: rewards 23.47 +/- 13.63
Iteration 295/500: rewards 21.66 +/- 11.3
Iteration 296/500: rewards 22.16 +/- 12.46
Iteration 297/500: rewards 22.94 +/- 12.22
Iteration 298/500: rewards 17.19 +/- 6.07
Iteration 299/500: rewards 17.47 +/- 9.12
Iteration 300/500: rewards 20.91 +/- 9.76
Iteration 301/500: rewards 22.22 +/- 12.66
Iteration 302/500: rewards 17.47 +/- 7.53
Iteration 303/500: rewards 20.38 +/- 10.51
Iteration 304/500: rewards 24.09 +/- 18.04
Iteration 305/500: rewards 19.56 +/- 12.85
Iteration 306/500: rewards 20.09 +/- 9.89
Iteration 307/500: rewards 18.56 +/- 8.59
Iteration 308/500: rewards 20.41 +/- 10.05
Iteration 309/500: rewards 20.5 +/- 11.85
Iteration 310/500: rewards 21.19 +/- 9.4
Iteration 311/500: rewards 17.72 +/- 6.6
Iteration 312/500: rewards 18.31 +/- 7.98
Iteration 313/500: rewards 18.78 +/- 8.0
Iteration 314/500: rewards 18.94 +/- 7.4
Iteration 315/500: rewards 22.59 +/- 11.43
Iteration 316/500: rewards 19.62 +/- 7.81
```

```
Iteration 317/500: rewards 20.81 +/- 10.33
Iteration 318/500: rewards 21.34 +/- 9.82
Iteration 319/500: rewards 19.0 +/- 6.73
Iteration 320/500: rewards 23.34 +/- 10.81
Iteration 321/500: rewards 16.16 +/- 4.95
Iteration 322/500: rewards 16.28 +/- 5.44
Iteration 323/500: rewards 19.75 +/- 9.8
Iteration 324/500: rewards 19.75 +/- 10.14
Iteration 325/500: rewards 18.84 +/- 8.44
Iteration 326/500: rewards 19.56 +/- 8.31
Iteration 327/500: rewards 19.5 +/- 7.67
Iteration 328/500: rewards 24.25 +/- 16.83
Iteration 329/500: rewards 23.06 +/- 15.6
Iteration 330/500: rewards 21.16 +/- 9.95
Iteration 331/500: rewards 21.94 +/- 13.81
Iteration 332/500: rewards 19.59 +/- 8.58
Iteration 333/500: rewards 19.16 +/- 8.88
Iteration 334/500: rewards 19.69 +/- 8.38
Iteration 335/500: rewards 19.53 +/- 9.94
Iteration 336/500: rewards 21.25 +/- 12.77
Iteration 337/500: rewards 18.53 +/- 10.84
Iteration 338/500: rewards 21.0 +/- 9.87
Iteration 339/500: rewards 18.56 +/- 7.18
Iteration 340/500: rewards 18.28 +/- 8.67
Iteration 341/500: rewards 18.34 +/- 8.27
Iteration 342/500: rewards 18.88 +/- 7.45
Iteration 343/500: rewards 22.12 +/- 14.29
Iteration 344/500: rewards 17.34 +/- 6.71
Iteration 345/500: rewards 19.91 +/- 11.06
Iteration 346/500: rewards 18.06 +/- 8.09
Iteration 347/500: rewards 20.66 +/- 9.83
Iteration 348/500: rewards 20.03 +/- 8.37
Iteration 349/500: rewards 18.31 +/- 5.9
Iteration 350/500: rewards 20.72 +/- 13.03
Iteration 351/500: rewards 19.62 +/- 11.14
Iteration 352/500: rewards 23.59 +/- 14.08
Iteration 353/500: rewards 20.09 +/- 8.98
Iteration 354/500: rewards 19.44 +/- 8.37
Iteration 355/500: rewards 17.59 +/- 8.39
Iteration 356/500: rewards 21.06 +/- 13.47
Iteration 357/500: rewards 19.69 +/- 10.91
Iteration 358/500: rewards 20.72 +/- 7.44
Iteration 359/500: rewards 21.41 +/- 9.34
Iteration 360/500: rewards 22.25 +/- 20.53
Iteration 361/500: rewards 19.28 +/- 10.48
Iteration 362/500: rewards 22.16 +/- 9.8
Iteration 363/500: rewards 18.34 +/- 7.12
Iteration 364/500: rewards 21.56 +/- 15.17
```

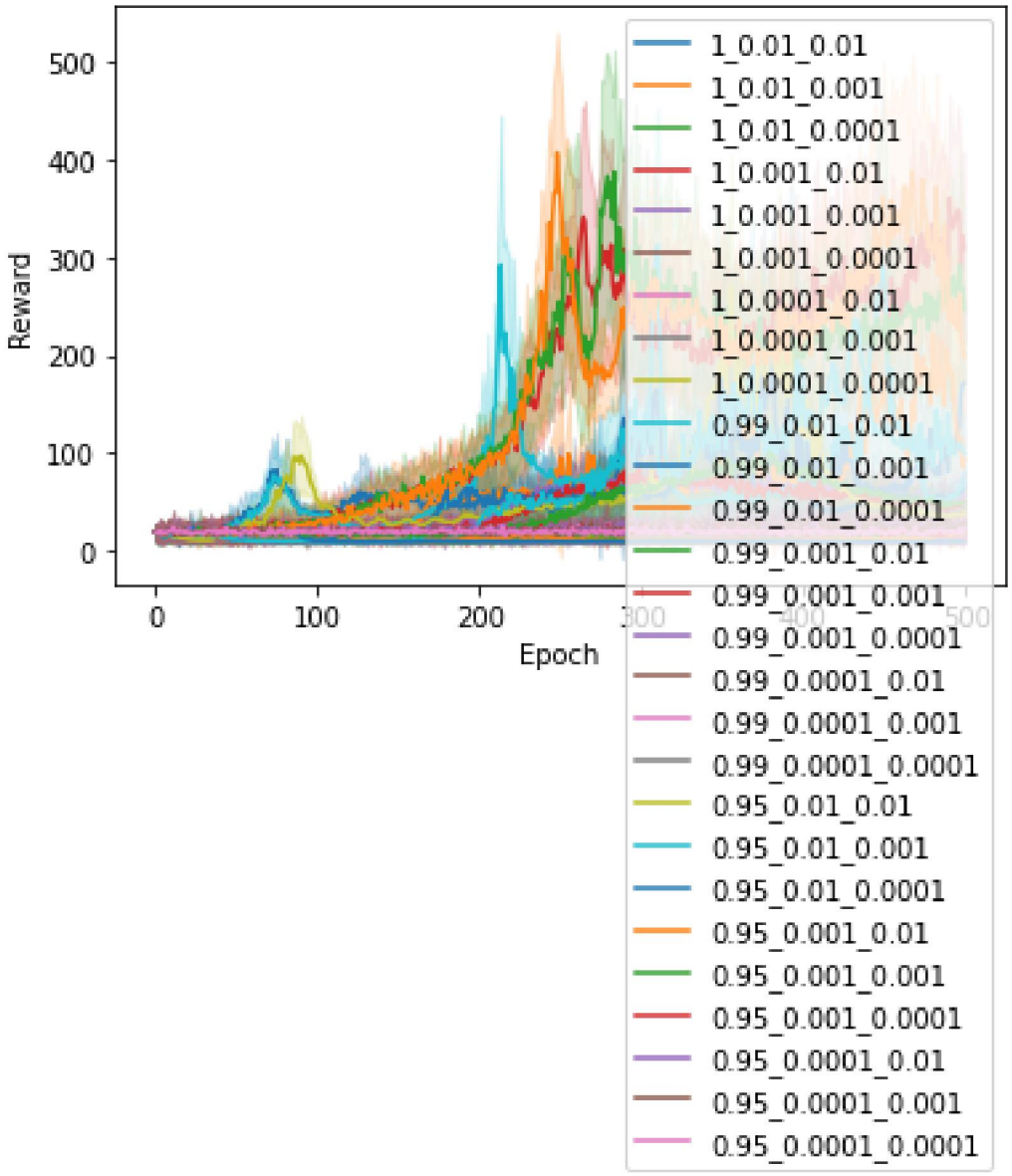
```
Iteration 365/500: rewards 20.22 +/- 12.23
Iteration 366/500: rewards 18.28 +/- 6.11
Iteration 367/500: rewards 21.19 +/- 7.93
Iteration 368/500: rewards 21.12 +/- 8.2
Iteration 369/500: rewards 19.38 +/- 8.63
Iteration 370/500: rewards 22.75 +/- 13.7
Iteration 371/500: rewards 19.78 +/- 10.27
Iteration 372/500: rewards 19.16 +/- 8.45
Iteration 373/500: rewards 17.28 +/- 8.22
Iteration 374/500: rewards 19.69 +/- 10.9
Iteration 375/500: rewards 20.78 +/- 10.57
Iteration 376/500: rewards 20.38 +/- 14.19
Iteration 377/500: rewards 19.59 +/- 12.13
Iteration 378/500: rewards 18.53 +/- 7.15
Iteration 379/500: rewards 18.34 +/- 8.34
Iteration 380/500: rewards 17.84 +/- 9.43
Iteration 381/500: rewards 19.78 +/- 8.55
Iteration 382/500: rewards 19.12 +/- 10.34
Iteration 383/500: rewards 19.06 +/- 12.58
Iteration 384/500: rewards 20.56 +/- 9.49
Iteration 385/500: rewards 20.0 +/- 10.86
Iteration 386/500: rewards 16.94 +/- 8.81
Iteration 387/500: rewards 23.41 +/- 17.53
Iteration 388/500: rewards 15.75 +/- 5.23
Iteration 389/500: rewards 18.91 +/- 7.76
Iteration 390/500: rewards 18.5 +/- 7.34
Iteration 391/500: rewards 18.09 +/- 7.88
Iteration 392/500: rewards 21.38 +/- 9.04
Iteration 393/500: rewards 20.75 +/- 9.0
Iteration 394/500: rewards 18.94 +/- 9.11
Iteration 395/500: rewards 19.91 +/- 10.92
Iteration 396/500: rewards 18.19 +/- 9.39
Iteration 397/500: rewards 21.0 +/- 10.02
Iteration 398/500: rewards 16.56 +/- 5.15
Iteration 399/500: rewards 18.12 +/- 7.25
Iteration 400/500: rewards 18.38 +/- 10.81
Iteration 401/500: rewards 17.88 +/- 7.26
Iteration 402/500: rewards 22.34 +/- 16.34
Iteration 403/500: rewards 20.56 +/- 12.09
Iteration 404/500: rewards 21.5 +/- 8.42
Iteration 405/500: rewards 17.5 +/- 7.2
Iteration 406/500: rewards 21.0 +/- 11.91
Iteration 407/500: rewards 19.44 +/- 11.27
Iteration 408/500: rewards 21.0 +/- 10.11
Iteration 409/500: rewards 21.06 +/- 12.06
Iteration 410/500: rewards 22.81 +/- 11.73
Iteration 411/500: rewards 17.59 +/- 7.56
Iteration 412/500: rewards 19.5 +/- 8.95
```

```
Iteration 413/500: rewards 17.31 +/- 6.91
Iteration 414/500: rewards 22.0 +/- 13.94
Iteration 415/500: rewards 18.91 +/- 6.49
Iteration 416/500: rewards 19.41 +/- 8.51
Iteration 417/500: rewards 19.0 +/- 8.79
Iteration 418/500: rewards 19.03 +/- 7.06
Iteration 419/500: rewards 19.44 +/- 9.02
Iteration 420/500: rewards 19.12 +/- 9.45
Iteration 421/500: rewards 20.47 +/- 10.54
Iteration 422/500: rewards 19.31 +/- 7.31
Iteration 423/500: rewards 18.28 +/- 8.21
Iteration 424/500: rewards 19.88 +/- 7.63
Iteration 425/500: rewards 20.34 +/- 10.68
Iteration 426/500: rewards 19.62 +/- 11.49
Iteration 427/500: rewards 19.41 +/- 7.74
Iteration 428/500: rewards 19.03 +/- 8.28
Iteration 429/500: rewards 15.41 +/- 5.71
Iteration 430/500: rewards 18.44 +/- 8.42
Iteration 431/500: rewards 19.66 +/- 8.55
Iteration 432/500: rewards 18.88 +/- 6.33
Iteration 433/500: rewards 16.84 +/- 7.48
Iteration 434/500: rewards 18.44 +/- 6.94
Iteration 435/500: rewards 22.66 +/- 14.15
Iteration 436/500: rewards 20.16 +/- 10.45
Iteration 437/500: rewards 18.59 +/- 6.3
Iteration 438/500: rewards 15.81 +/- 5.81
Iteration 439/500: rewards 19.84 +/- 8.53
Iteration 440/500: rewards 18.5 +/- 7.3
Iteration 441/500: rewards 16.94 +/- 8.36
Iteration 442/500: rewards 21.28 +/- 9.76
Iteration 443/500: rewards 23.41 +/- 14.12
Iteration 444/500: rewards 18.5 +/- 8.07
Iteration 445/500: rewards 22.12 +/- 10.38
Iteration 446/500: rewards 21.47 +/- 10.72
Iteration 447/500: rewards 18.47 +/- 9.51
Iteration 448/500: rewards 19.56 +/- 11.7
Iteration 449/500: rewards 19.81 +/- 10.95
Iteration 450/500: rewards 18.12 +/- 6.64
Iteration 451/500: rewards 18.47 +/- 12.28
Iteration 452/500: rewards 16.09 +/- 5.9
Iteration 453/500: rewards 18.06 +/- 9.81
Iteration 454/500: rewards 21.94 +/- 11.75
Iteration 455/500: rewards 18.44 +/- 6.53
Iteration 456/500: rewards 17.28 +/- 7.51
Iteration 457/500: rewards 20.03 +/- 11.23
Iteration 458/500: rewards 22.03 +/- 9.47
Iteration 459/500: rewards 16.97 +/- 7.87
Iteration 460/500: rewards 18.75 +/- 7.83
```

```
Iteration 461/500: rewards 23.44 +/- 12.68
Iteration 462/500: rewards 22.03 +/- 16.24
Iteration 463/500: rewards 21.16 +/- 10.05
Iteration 464/500: rewards 20.25 +/- 9.98
Iteration 465/500: rewards 17.44 +/- 7.91
Iteration 466/500: rewards 17.72 +/- 6.48
Iteration 467/500: rewards 17.97 +/- 9.69
Iteration 468/500: rewards 17.91 +/- 6.54
Iteration 469/500: rewards 19.44 +/- 8.81
Iteration 470/500: rewards 25.41 +/- 15.95
Iteration 471/500: rewards 18.38 +/- 7.35
Iteration 472/500: rewards 16.94 +/- 6.63
Iteration 473/500: rewards 16.53 +/- 6.56
Iteration 474/500: rewards 17.44 +/- 6.12
Iteration 475/500: rewards 15.5 +/- 5.11
Iteration 476/500: rewards 18.69 +/- 11.56
Iteration 477/500: rewards 20.91 +/- 12.3
Iteration 478/500: rewards 18.97 +/- 8.63
Iteration 479/500: rewards 17.47 +/- 6.8
Iteration 480/500: rewards 19.62 +/- 10.38
Iteration 481/500: rewards 20.41 +/- 11.55
Iteration 482/500: rewards 19.0 +/- 7.25
Iteration 483/500: rewards 18.59 +/- 9.52
Iteration 484/500: rewards 21.56 +/- 12.77
Iteration 485/500: rewards 21.88 +/- 10.72
Iteration 486/500: rewards 20.12 +/- 9.95
Iteration 487/500: rewards 19.12 +/- 7.71
Iteration 488/500: rewards 17.56 +/- 7.95
Iteration 489/500: rewards 20.06 +/- 10.28
Iteration 490/500: rewards 20.56 +/- 10.02
Iteration 491/500: rewards 20.22 +/- 9.65
Iteration 492/500: rewards 17.06 +/- 6.05
Iteration 493/500: rewards 20.72 +/- 13.94
Iteration 494/500: rewards 19.0 +/- 9.21
Iteration 495/500: rewards 17.88 +/- 6.87
Iteration 496/500: rewards 18.38 +/- 7.98
Iteration 497/500: rewards 23.34 +/- 12.63
Iteration 498/500: rewards 19.22 +/- 11.95
Iteration 499/500: rewards 16.59 +/- 7.24
Iteration 500/500: rewards 21.88 +/- 14.32
The average reward is 19.288125
```



[ ]:
```
plot_everything(rewards)
```



I am sorry if the legend is overlaying the figure. My session crahed and I didn't have enough time to rerun the entire gridsearch

[ ]:
```
def find_max (rewards):
  ks = list(rewards.keys())
  max = -np.Inf
  key=None
  for k in ks:
    if rewards[k] >max:
      max = rewards[k]
      key = k
```

```
  return key, max

key, max = find_max (avg_rewards)
```

[ ]:
```
best_gamma = float(key.split('_')[0])
best_value_lr = float(key.split('_')[-1])
best_policy_lr = float(key.split('_')[1])
```

[ ]:
```
print('The best gamma is: {}'.format(best_gamma))
print('The best value lr is: {}'.format(best_value_lr))
print('The best policy lr is: {}'.format(best_policy_lr))
print('The avg reward associated with the best config is: {}'.format(max))
```

```
The best gamma is: 0.95
The best value lr is: 0.01
The best policy lr is: 0.001
The avg reward associated with the best config is: 310.58125
```

[ ]:
```
# Provide your best config here and run this cell
config = {
    'env_id': 'CartPole-v1',
    'seed': 8953,
    'gamma': 0.95,
    'policy_layers': [16, 8],
    'policy_learning_rate': 1e-3,
    'use_baseline': True,
    'value_layers': [30],
    'value_learning_rate': 1e-2,
}
agent = ActorCriticAgent(config)
ActorCritic_rewards2 = agent.train(n_episodes=100, n_iterations=500)
```

```
the device is: cpu
Iteration 1/500: rewards 18.37 +/- 8.67
Iteration 2/500: rewards 19.42 +/- 9.88
Iteration 3/500: rewards 20.06 +/- 10.04
Iteration 4/500: rewards 19.79 +/- 10.07
Iteration 5/500: rewards 19.47 +/- 9.47
Iteration 6/500: rewards 18.77 +/- 9.11
Iteration 7/500: rewards 20.47 +/- 11.08
Iteration 8/500: rewards 17.76 +/- 8.03
Iteration 9/500: rewards 19.41 +/- 10.83
Iteration 10/500: rewards 20.19 +/- 8.8
Iteration 11/500: rewards 20.27 +/- 10.63
Iteration 12/500: rewards 20.26 +/- 11.0
Iteration 13/500: rewards 19.26 +/- 10.03
Iteration 14/500: rewards 20.96 +/- 10.69
Iteration 15/500: rewards 19.53 +/- 9.36
```

```
Iteration 16/500: rewards 20.52 +/- 12.62
Iteration 17/500: rewards 18.61 +/- 8.22
Iteration 18/500: rewards 22.64 +/- 11.77
Iteration 19/500: rewards 20.14 +/- 12.22
Iteration 20/500: rewards 20.15 +/- 11.4
Iteration 21/500: rewards 18.41 +/- 9.29
Iteration 22/500: rewards 19.01 +/- 10.28
Iteration 23/500: rewards 18.87 +/- 9.14
Iteration 24/500: rewards 18.4 +/- 8.89
Iteration 25/500: rewards 21.01 +/- 10.41
Iteration 26/500: rewards 19.81 +/- 9.75
Iteration 27/500: rewards 18.88 +/- 7.78
Iteration 28/500: rewards 19.47 +/- 8.55
Iteration 29/500: rewards 20.42 +/- 10.39
Iteration 30/500: rewards 20.85 +/- 12.28
Iteration 31/500: rewards 20.57 +/- 11.25
Iteration 32/500: rewards 19.99 +/- 9.91
Iteration 33/500: rewards 20.08 +/- 8.64
Iteration 34/500: rewards 20.64 +/- 10.82
Iteration 35/500: rewards 18.92 +/- 8.71
Iteration 36/500: rewards 19.42 +/- 8.56
Iteration 37/500: rewards 19.69 +/- 10.73
Iteration 38/500: rewards 19.32 +/- 9.47
Iteration 39/500: rewards 19.39 +/- 9.18
Iteration 40/500: rewards 22.69 +/- 14.48
Iteration 41/500: rewards 19.9 +/- 11.08
Iteration 42/500: rewards 21.6 +/- 11.75
Iteration 43/500: rewards 21.03 +/- 10.55
Iteration 44/500: rewards 20.46 +/- 10.32
Iteration 45/500: rewards 20.64 +/- 11.41
Iteration 46/500: rewards 21.15 +/- 9.69
Iteration 47/500: rewards 21.12 +/- 11.43
Iteration 48/500: rewards 21.14 +/- 11.71
Iteration 49/500: rewards 21.43 +/- 11.65
Iteration 50/500: rewards 20.45 +/- 9.04
Iteration 51/500: rewards 23.71 +/- 12.56
Iteration 52/500: rewards 21.01 +/- 11.94
Iteration 53/500: rewards 23.02 +/- 12.62
Iteration 54/500: rewards 22.45 +/- 12.88
Iteration 55/500: rewards 22.09 +/- 9.66
Iteration 56/500: rewards 22.3 +/- 12.29
Iteration 57/500: rewards 24.84 +/- 13.63
Iteration 58/500: rewards 23.88 +/- 11.4
Iteration 59/500: rewards 25.19 +/- 13.24
Iteration 60/500: rewards 23.82 +/- 11.91
Iteration 61/500: rewards 22.5 +/- 10.59
Iteration 62/500: rewards 23.92 +/- 11.3
Iteration 63/500: rewards 24.13 +/- 12.62
```

```
Iteration 64/500: rewards 25.4 +/- 13.81
Iteration 65/500: rewards 26.39 +/- 16.51
Iteration 66/500: rewards 22.1 +/- 12.0
Iteration 67/500: rewards 23.19 +/- 11.56
Iteration 68/500: rewards 26.0 +/- 12.48
Iteration 69/500: rewards 24.57 +/- 12.62
Iteration 70/500: rewards 26.36 +/- 13.88
Iteration 71/500: rewards 26.07 +/- 14.83
Iteration 72/500: rewards 23.49 +/- 10.73
Iteration 73/500: rewards 26.59 +/- 15.93
Iteration 74/500: rewards 26.79 +/- 15.32
Iteration 75/500: rewards 24.6 +/- 12.0
Iteration 76/500: rewards 27.29 +/- 15.39
Iteration 77/500: rewards 27.49 +/- 16.29
Iteration 78/500: rewards 28.26 +/- 22.09
Iteration 79/500: rewards 28.86 +/- 15.95
Iteration 80/500: rewards 28.83 +/- 17.87
Iteration 81/500: rewards 30.37 +/- 15.94
Iteration 82/500: rewards 28.04 +/- 15.66
Iteration 83/500: rewards 29.16 +/- 14.46
Iteration 84/500: rewards 32.54 +/- 19.01
Iteration 85/500: rewards 29.13 +/- 16.42
Iteration 86/500: rewards 31.85 +/- 18.15
Iteration 87/500: rewards 31.52 +/- 16.35
Iteration 88/500: rewards 29.23 +/- 17.19
Iteration 89/500: rewards 32.16 +/- 16.74
Iteration 90/500: rewards 29.84 +/- 16.69
Iteration 91/500: rewards 33.18 +/- 19.8
Iteration 92/500: rewards 32.53 +/- 16.75
Iteration 93/500: rewards 30.84 +/- 18.71
Iteration 94/500: rewards 32.66 +/- 18.74
Iteration 95/500: rewards 35.34 +/- 21.27
Iteration 96/500: rewards 29.41 +/- 13.34
Iteration 97/500: rewards 34.53 +/- 21.26
Iteration 98/500: rewards 35.22 +/- 18.83
Iteration 99/500: rewards 34.56 +/- 16.35
Iteration 100/500: rewards 36.57 +/- 20.96
Iteration 101/500: rewards 34.08 +/- 18.53
Iteration 102/500: rewards 36.09 +/- 21.47
Iteration 103/500: rewards 39.87 +/- 25.75
Iteration 104/500: rewards 34.44 +/- 20.11
Iteration 105/500: rewards 37.38 +/- 18.2
Iteration 106/500: rewards 39.0 +/- 21.29
Iteration 107/500: rewards 34.04 +/- 19.74
Iteration 108/500: rewards 40.35 +/- 22.37
Iteration 109/500: rewards 38.34 +/- 18.27
Iteration 110/500: rewards 40.71 +/- 25.18
Iteration 111/500: rewards 40.83 +/- 23.33
```

```
Iteration 112/500: rewards 39.74 +/- 23.37
Iteration 113/500: rewards 38.33 +/- 20.12
Iteration 114/500: rewards 42.2 +/- 23.1
Iteration 115/500: rewards 38.39 +/- 19.34
Iteration 116/500: rewards 38.74 +/- 21.84
Iteration 117/500: rewards 42.59 +/- 19.79
Iteration 118/500: rewards 43.5 +/- 22.02
Iteration 119/500: rewards 43.39 +/- 25.33
Iteration 120/500: rewards 42.2 +/- 22.31
Iteration 121/500: rewards 44.11 +/- 22.59
Iteration 122/500: rewards 45.61 +/- 27.27
Iteration 123/500: rewards 45.98 +/- 22.32
Iteration 124/500: rewards 47.97 +/- 25.97
Iteration 125/500: rewards 42.68 +/- 20.3
Iteration 126/500: rewards 47.37 +/- 22.25
Iteration 127/500: rewards 43.03 +/- 21.24
Iteration 128/500: rewards 45.59 +/- 23.96
Iteration 129/500: rewards 42.95 +/- 22.24
Iteration 130/500: rewards 44.47 +/- 24.19
Iteration 131/500: rewards 49.39 +/- 29.49
Iteration 132/500: rewards 52.49 +/- 26.46
Iteration 133/500: rewards 46.69 +/- 24.54
Iteration 134/500: rewards 46.87 +/- 21.69
Iteration 135/500: rewards 51.02 +/- 27.86
Iteration 136/500: rewards 50.83 +/- 27.39
Iteration 137/500: rewards 52.31 +/- 28.16
Iteration 138/500: rewards 49.96 +/- 22.23
Iteration 139/500: rewards 57.42 +/- 26.9
Iteration 140/500: rewards 55.21 +/- 30.38
Iteration 141/500: rewards 53.32 +/- 27.35
Iteration 142/500: rewards 56.0 +/- 25.5
Iteration 143/500: rewards 53.19 +/- 21.19
Iteration 144/500: rewards 57.54 +/- 22.75
Iteration 145/500: rewards 53.23 +/- 27.41
Iteration 146/500: rewards 63.42 +/- 27.64
Iteration 147/500: rewards 60.33 +/- 30.51
Iteration 148/500: rewards 57.7 +/- 24.65
Iteration 149/500: rewards 59.29 +/- 31.02
Iteration 150/500: rewards 64.22 +/- 27.51
Iteration 151/500: rewards 63.69 +/- 27.45
Iteration 152/500: rewards 60.91 +/- 25.77
Iteration 153/500: rewards 60.97 +/- 27.55
Iteration 154/500: rewards 64.94 +/- 35.56
Iteration 155/500: rewards 63.83 +/- 27.32
Iteration 156/500: rewards 60.05 +/- 24.98
Iteration 157/500: rewards 67.12 +/- 29.0
Iteration 158/500: rewards 70.97 +/- 30.67
Iteration 159/500: rewards 70.13 +/- 33.94
```

```
Iteration 160/500: rewards 64.44 +/- 27.02
Iteration 161/500: rewards 70.11 +/- 31.09
Iteration 162/500: rewards 72.0 +/- 24.45
Iteration 163/500: rewards 71.21 +/- 30.11
Iteration 164/500: rewards 73.72 +/- 32.49
Iteration 165/500: rewards 72.04 +/- 32.68
Iteration 166/500: rewards 75.98 +/- 39.23
Iteration 167/500: rewards 78.53 +/- 40.2
Iteration 168/500: rewards 70.25 +/- 30.13
Iteration 169/500: rewards 81.35 +/- 38.46
Iteration 170/500: rewards 85.65 +/- 40.07
Iteration 171/500: rewards 79.14 +/- 35.17
Iteration 172/500: rewards 89.13 +/- 35.47
Iteration 173/500: rewards 80.9 +/- 36.37
Iteration 174/500: rewards 90.22 +/- 43.07
Iteration 175/500: rewards 88.54 +/- 36.35
Iteration 176/500: rewards 90.48 +/- 36.25
Iteration 177/500: rewards 91.62 +/- 37.56
Iteration 178/500: rewards 95.56 +/- 42.9
Iteration 179/500: rewards 95.61 +/- 42.27
Iteration 180/500: rewards 94.34 +/- 33.25
Iteration 181/500: rewards 101.72 +/- 49.08
Iteration 182/500: rewards 100.29 +/- 47.57
Iteration 183/500: rewards 108.89 +/- 48.9
Iteration 184/500: rewards 107.83 +/- 46.69
Iteration 185/500: rewards 119.11 +/- 55.47
Iteration 186/500: rewards 112.88 +/- 49.06
Iteration 187/500: rewards 107.89 +/- 39.33
Iteration 188/500: rewards 120.36 +/- 58.53
Iteration 189/500: rewards 119.43 +/- 58.37
Iteration 190/500: rewards 122.12 +/- 47.81
Iteration 191/500: rewards 127.98 +/- 60.57
Iteration 192/500: rewards 125.94 +/- 57.28
Iteration 193/500: rewards 132.0 +/- 54.52
Iteration 194/500: rewards 138.98 +/- 52.86
Iteration 195/500: rewards 139.87 +/- 58.34
Iteration 196/500: rewards 144.71 +/- 58.23
Iteration 197/500: rewards 136.56 +/- 57.5
Iteration 198/500: rewards 155.71 +/- 57.42
Iteration 199/500: rewards 169.21 +/- 71.41
Iteration 200/500: rewards 162.6 +/- 65.23
Iteration 201/500: rewards 167.8 +/- 67.2
Iteration 202/500: rewards 178.41 +/- 61.61
Iteration 203/500: rewards 187.03 +/- 83.36
Iteration 204/500: rewards 174.28 +/- 77.21
Iteration 205/500: rewards 195.77 +/- 83.54
Iteration 206/500: rewards 187.26 +/- 77.83
Iteration 207/500: rewards 182.63 +/- 68.5
```

```
Iteration 208/500: rewards 183.01 +/- 71.73
Iteration 209/500: rewards 210.61 +/- 79.56
Iteration 210/500: rewards 222.1 +/- 80.77
Iteration 211/500: rewards 201.84 +/- 81.83
Iteration 212/500: rewards 205.11 +/- 86.2
Iteration 213/500: rewards 222.18 +/- 90.22
Iteration 214/500: rewards 238.22 +/- 95.14
Iteration 215/500: rewards 238.14 +/- 96.01
Iteration 216/500: rewards 239.48 +/- 99.76
Iteration 217/500: rewards 254.63 +/- 116.12
Iteration 218/500: rewards 240.72 +/- 97.69
Iteration 219/500: rewards 238.04 +/- 97.24
Iteration 220/500: rewards 233.28 +/- 98.02
Iteration 221/500: rewards 238.25 +/- 96.39
Iteration 222/500: rewards 242.6 +/- 93.4
Iteration 223/500: rewards 244.91 +/- 105.51
Iteration 224/500: rewards 225.79 +/- 76.41
Iteration 225/500: rewards 257.22 +/- 98.68
Iteration 226/500: rewards 251.93 +/- 92.75
Iteration 227/500: rewards 246.39 +/- 92.1
Iteration 228/500: rewards 250.82 +/- 98.1
Iteration 229/500: rewards 245.41 +/- 79.39
Iteration 230/500: rewards 256.72 +/- 96.09
Iteration 231/500: rewards 243.6 +/- 92.36
Iteration 232/500: rewards 244.15 +/- 86.18
Iteration 233/500: rewards 221.77 +/- 72.05
Iteration 234/500: rewards 236.04 +/- 80.64
Iteration 235/500: rewards 218.12 +/- 59.96
Iteration 236/500: rewards 247.32 +/- 92.07
Iteration 237/500: rewards 239.26 +/- 82.91
Iteration 238/500: rewards 233.85 +/- 83.84
Iteration 239/500: rewards 245.08 +/- 90.51
Iteration 240/500: rewards 233.42 +/- 84.71
Iteration 241/500: rewards 242.43 +/- 91.27
Iteration 242/500: rewards 222.45 +/- 85.44
Iteration 243/500: rewards 223.12 +/- 74.74
Iteration 244/500: rewards 225.4 +/- 77.14
Iteration 245/500: rewards 227.38 +/- 77.31
Iteration 246/500: rewards 228.07 +/- 76.71
Iteration 247/500: rewards 237.32 +/- 85.08
Iteration 248/500: rewards 239.07 +/- 83.14
Iteration 249/500: rewards 233.7 +/- 89.67
Iteration 250/500: rewards 245.88 +/- 82.37
Iteration 251/500: rewards 245.21 +/- 97.57
Iteration 252/500: rewards 236.61 +/- 72.11
Iteration 253/500: rewards 223.2 +/- 74.72
Iteration 254/500: rewards 225.24 +/- 76.64
Iteration 255/500: rewards 239.92 +/- 81.31
```

```
Iteration 256/500: rewards 238.75 +/- 85.61
Iteration 257/500: rewards 245.73 +/- 79.4
Iteration 258/500: rewards 222.01 +/- 80.63
Iteration 259/500: rewards 220.85 +/- 69.24
Iteration 260/500: rewards 232.98 +/- 71.11
Iteration 261/500: rewards 229.65 +/- 64.48
Iteration 262/500: rewards 225.48 +/- 77.33
Iteration 263/500: rewards 223.29 +/- 76.58
Iteration 264/500: rewards 232.88 +/- 78.51
Iteration 265/500: rewards 236.05 +/- 85.51
Iteration 266/500: rewards 238.87 +/- 82.58
Iteration 267/500: rewards 229.0 +/- 81.97
Iteration 268/500: rewards 236.67 +/- 83.51
Iteration 269/500: rewards 240.4 +/- 82.65
Iteration 270/500: rewards 231.33 +/- 71.36
Iteration 271/500: rewards 251.46 +/- 95.72
Iteration 272/500: rewards 252.58 +/- 91.83
Iteration 273/500: rewards 234.4 +/- 90.83
Iteration 274/500: rewards 227.82 +/- 80.9
Iteration 275/500: rewards 236.79 +/- 78.24
Iteration 276/500: rewards 238.2 +/- 72.77
Iteration 277/500: rewards 235.58 +/- 91.69
Iteration 278/500: rewards 236.25 +/- 89.49
Iteration 279/500: rewards 237.78 +/- 86.77
Iteration 280/500: rewards 256.43 +/- 82.38
Iteration 281/500: rewards 251.76 +/- 87.72
Iteration 282/500: rewards 245.64 +/- 83.78
Iteration 283/500: rewards 235.78 +/- 83.15
Iteration 284/500: rewards 239.93 +/- 79.86
Iteration 285/500: rewards 240.7 +/- 82.09
Iteration 286/500: rewards 222.14 +/- 77.71
Iteration 287/500: rewards 232.52 +/- 82.66
Iteration 288/500: rewards 245.05 +/- 92.21
Iteration 289/500: rewards 241.6 +/- 82.19
Iteration 290/500: rewards 248.89 +/- 90.45
Iteration 291/500: rewards 239.34 +/- 75.72
Iteration 292/500: rewards 239.01 +/- 84.49
Iteration 293/500: rewards 240.22 +/- 79.87
Iteration 294/500: rewards 240.53 +/- 90.58
Iteration 295/500: rewards 229.18 +/- 67.35
Iteration 296/500: rewards 231.94 +/- 84.6
Iteration 297/500: rewards 233.77 +/- 76.56
Iteration 298/500: rewards 224.26 +/- 65.85
Iteration 299/500: rewards 217.54 +/- 63.0
Iteration 300/500: rewards 239.03 +/- 85.63
Iteration 301/500: rewards 241.63 +/- 83.83
Iteration 302/500: rewards 241.37 +/- 85.74
Iteration 303/500: rewards 225.45 +/- 77.97
```

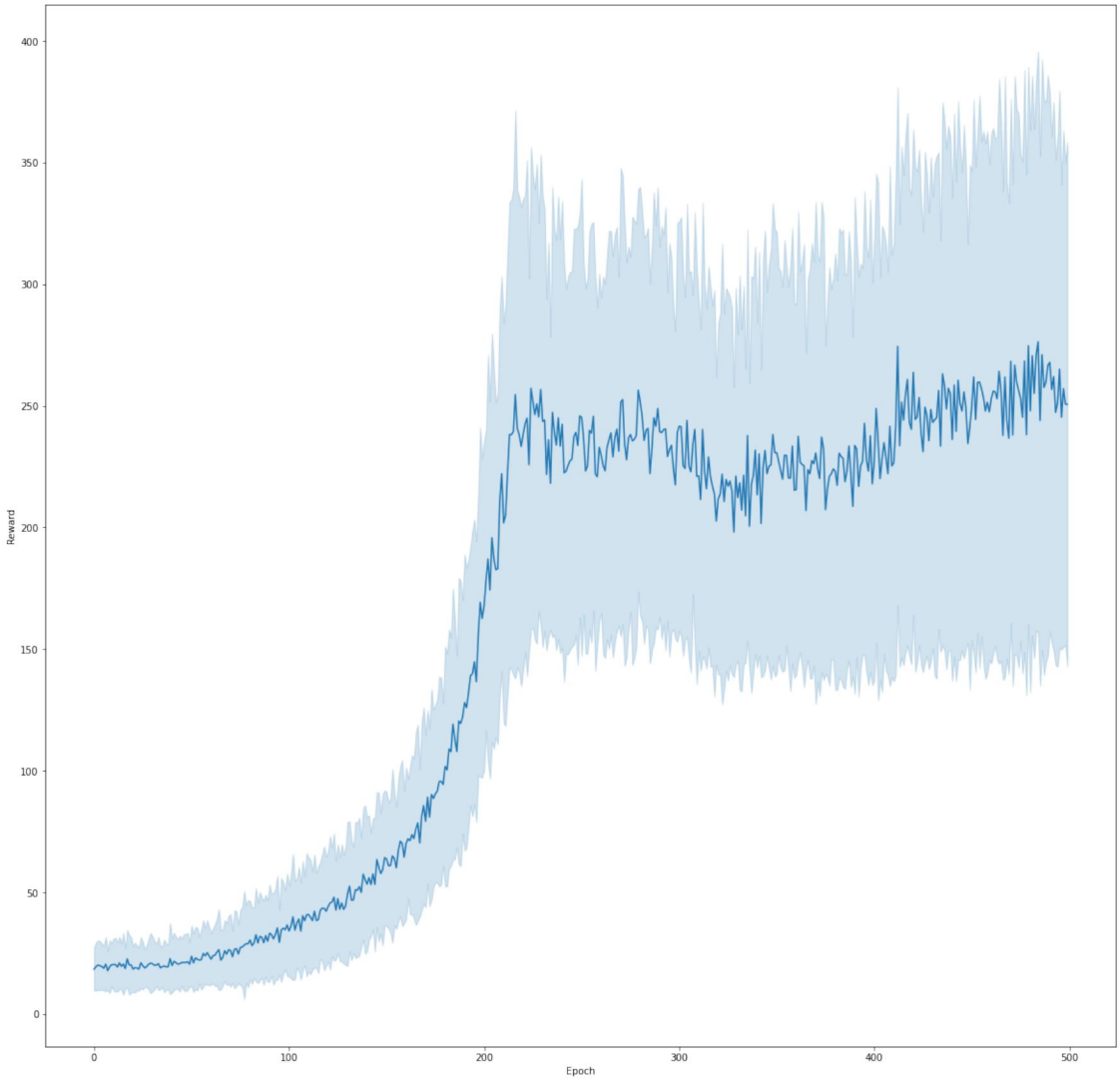
```
Iteration 304/500: rewards 224.24 +/- 70.18
Iteration 305/500: rewards 244.0 +/- 88.66
Iteration 306/500: rewards 225.2 +/- 79.06
Iteration 307/500: rewards 222.86 +/- 82.11
Iteration 308/500: rewards 234.11 +/- 61.09
Iteration 309/500: rewards 240.3 +/- 88.77
Iteration 310/500: rewards 221.04 +/- 85.1
Iteration 311/500: rewards 221.2 +/- 71.67
Iteration 312/500: rewards 211.47 +/- 70.05
Iteration 313/500: rewards 240.31 +/- 92.82
Iteration 314/500: rewards 222.64 +/- 76.92
Iteration 315/500: rewards 215.87 +/- 73.64
Iteration 316/500: rewards 229.0 +/- 77.98
Iteration 317/500: rewards 221.03 +/- 77.98
Iteration 318/500: rewards 217.13 +/- 73.64
Iteration 319/500: rewards 214.01 +/- 83.06
Iteration 320/500: rewards 202.61 +/- 58.8
Iteration 321/500: rewards 211.79 +/- 71.8
Iteration 322/500: rewards 213.72 +/- 73.75
Iteration 323/500: rewards 221.91 +/- 94.09
Iteration 324/500: rewards 210.66 +/- 76.99
Iteration 325/500: rewards 219.76 +/- 78.19
Iteration 326/500: rewards 216.91 +/- 79.49
Iteration 327/500: rewards 219.09 +/- 75.34
Iteration 328/500: rewards 215.11 +/- 75.02
Iteration 329/500: rewards 198.1 +/- 59.3
Iteration 330/500: rewards 221.12 +/- 76.94
Iteration 331/500: rewards 212.29 +/- 66.84
Iteration 332/500: rewards 218.35 +/- 84.56
Iteration 333/500: rewards 207.08 +/- 74.14
Iteration 334/500: rewards 221.48 +/- 77.29
Iteration 335/500: rewards 204.79 +/- 60.18
Iteration 336/500: rewards 237.79 +/- 84.1
Iteration 337/500: rewards 200.49 +/- 58.54
Iteration 338/500: rewards 217.65 +/- 85.23
Iteration 339/500: rewards 221.45 +/- 80.54
Iteration 340/500: rewards 231.78 +/- 83.26
Iteration 341/500: rewards 213.43 +/- 70.57
Iteration 342/500: rewards 230.24 +/- 82.53
Iteration 343/500: rewards 201.65 +/- 62.94
Iteration 344/500: rewards 224.68 +/- 85.25
Iteration 345/500: rewards 231.66 +/- 89.9
Iteration 346/500: rewards 222.09 +/- 74.38
Iteration 347/500: rewards 225.05 +/- 82.49
Iteration 348/500: rewards 225.69 +/- 87.73
Iteration 349/500: rewards 238.19 +/- 94.54
Iteration 350/500: rewards 230.7 +/- 91.52
Iteration 351/500: rewards 230.74 +/- 90.35
```

```
Iteration 352/500: rewards 227.14 +/- 79.17
Iteration 353/500: rewards 223.72 +/- 81.2
Iteration 354/500: rewards 219.87 +/- 78.78
Iteration 355/500: rewards 229.7 +/- 87.98
Iteration 356/500: rewards 229.61 +/- 77.53
Iteration 357/500: rewards 220.37 +/- 78.3
Iteration 358/500: rewards 220.28 +/- 87.15
Iteration 359/500: rewards 233.43 +/- 89.09
Iteration 360/500: rewards 215.29 +/- 76.79
Iteration 361/500: rewards 215.59 +/- 75.76
Iteration 362/500: rewards 237.45 +/- 91.8
Iteration 363/500: rewards 226.83 +/- 77.86
Iteration 364/500: rewards 225.87 +/- 83.11
Iteration 365/500: rewards 225.38 +/- 90.59
Iteration 366/500: rewards 206.92 +/- 64.72
Iteration 367/500: rewards 223.8 +/- 77.95
Iteration 368/500: rewards 222.09 +/- 83.8
Iteration 369/500: rewards 227.47 +/- 88.98
Iteration 370/500: rewards 226.24 +/- 82.72
Iteration 371/500: rewards 230.62 +/- 102.74
Iteration 372/500: rewards 224.16 +/- 87.21
Iteration 373/500: rewards 220.12 +/- 88.85
Iteration 374/500: rewards 237.16 +/- 96.17
Iteration 375/500: rewards 232.39 +/- 94.7
Iteration 376/500: rewards 207.29 +/- 67.34
Iteration 377/500: rewards 215.95 +/- 77.12
Iteration 378/500: rewards 221.0 +/- 85.57
Iteration 379/500: rewards 222.07 +/- 75.52
Iteration 380/500: rewards 224.07 +/- 78.2
Iteration 381/500: rewards 223.05 +/- 89.17
Iteration 382/500: rewards 217.29 +/- 83.77
Iteration 383/500: rewards 230.54 +/- 92.06
Iteration 384/500: rewards 229.16 +/- 91.38
Iteration 385/500: rewards 228.41 +/- 93.54
Iteration 386/500: rewards 218.82 +/- 84.68
Iteration 387/500: rewards 223.19 +/- 81.03
Iteration 388/500: rewards 233.46 +/- 87.82
Iteration 389/500: rewards 222.14 +/- 87.32
Iteration 390/500: rewards 208.64 +/- 69.5
Iteration 391/500: rewards 233.71 +/- 101.54
Iteration 392/500: rewards 232.23 +/- 85.13
Iteration 393/500: rewards 216.84 +/- 85.84
Iteration 394/500: rewards 225.2 +/- 82.83
Iteration 395/500: rewards 227.22 +/- 79.09
Iteration 396/500: rewards 242.82 +/- 94.89
Iteration 397/500: rewards 227.76 +/- 86.88
Iteration 398/500: rewards 223.16 +/- 87.53
Iteration 399/500: rewards 237.7 +/- 96.51
```

```
Iteration 400/500: rewards 217.93 +/- 82.39
Iteration 401/500: rewards 227.57 +/- 90.1
Iteration 402/500: rewards 248.93 +/- 95.98
Iteration 403/500: rewards 235.26 +/- 105.79
Iteration 404/500: rewards 220.08 +/- 82.35
Iteration 405/500: rewards 228.14 +/- 95.3
Iteration 406/500: rewards 234.84 +/- 86.51
Iteration 407/500: rewards 228.9 +/- 87.15
Iteration 408/500: rewards 222.11 +/- 82.63
Iteration 409/500: rewards 241.62 +/- 106.23
Iteration 410/500: rewards 225.29 +/- 86.38
Iteration 411/500: rewards 226.53 +/- 89.32
Iteration 412/500: rewards 241.45 +/- 101.7
Iteration 413/500: rewards 274.43 +/- 105.94
Iteration 414/500: rewards 233.58 +/- 90.59
Iteration 415/500: rewards 251.54 +/- 104.48
Iteration 416/500: rewards 244.12 +/- 100.27
Iteration 417/500: rewards 254.49 +/- 105.43
Iteration 418/500: rewards 260.85 +/- 108.97
Iteration 419/500: rewards 243.29 +/- 95.78
Iteration 420/500: rewards 240.32 +/- 96.1
Iteration 421/500: rewards 263.8 +/- 99.27
Iteration 422/500: rewards 244.39 +/- 103.23
Iteration 423/500: rewards 245.49 +/- 100.52
Iteration 424/500: rewards 253.44 +/- 101.35
Iteration 425/500: rewards 238.85 +/- 94.88
Iteration 426/500: rewards 231.17 +/- 90.21
Iteration 427/500: rewards 249.48 +/- 104.33
Iteration 428/500: rewards 245.97 +/- 98.32
Iteration 429/500: rewards 235.63 +/- 93.4
Iteration 430/500: rewards 248.5 +/- 103.26
Iteration 431/500: rewards 243.18 +/- 92.37
Iteration 432/500: rewards 244.24 +/- 104.72
Iteration 433/500: rewards 245.19 +/- 106.82
Iteration 434/500: rewards 256.27 +/- 97.49
Iteration 435/500: rewards 233.43 +/- 84.24
Iteration 436/500: rewards 263.14 +/- 111.12
Iteration 437/500: rewards 258.38 +/- 108.37
Iteration 438/500: rewards 248.7 +/- 106.43
Iteration 439/500: rewards 257.21 +/- 107.37
Iteration 440/500: rewards 255.04 +/- 105.0
Iteration 441/500: rewards 236.17 +/- 98.79
Iteration 442/500: rewards 258.51 +/- 111.16
Iteration 443/500: rewards 239.51 +/- 102.07
Iteration 444/500: rewards 260.52 +/- 114.18
Iteration 445/500: rewards 250.96 +/- 104.93
Iteration 446/500: rewards 247.86 +/- 97.89
Iteration 447/500: rewards 255.71 +/- 109.06
```

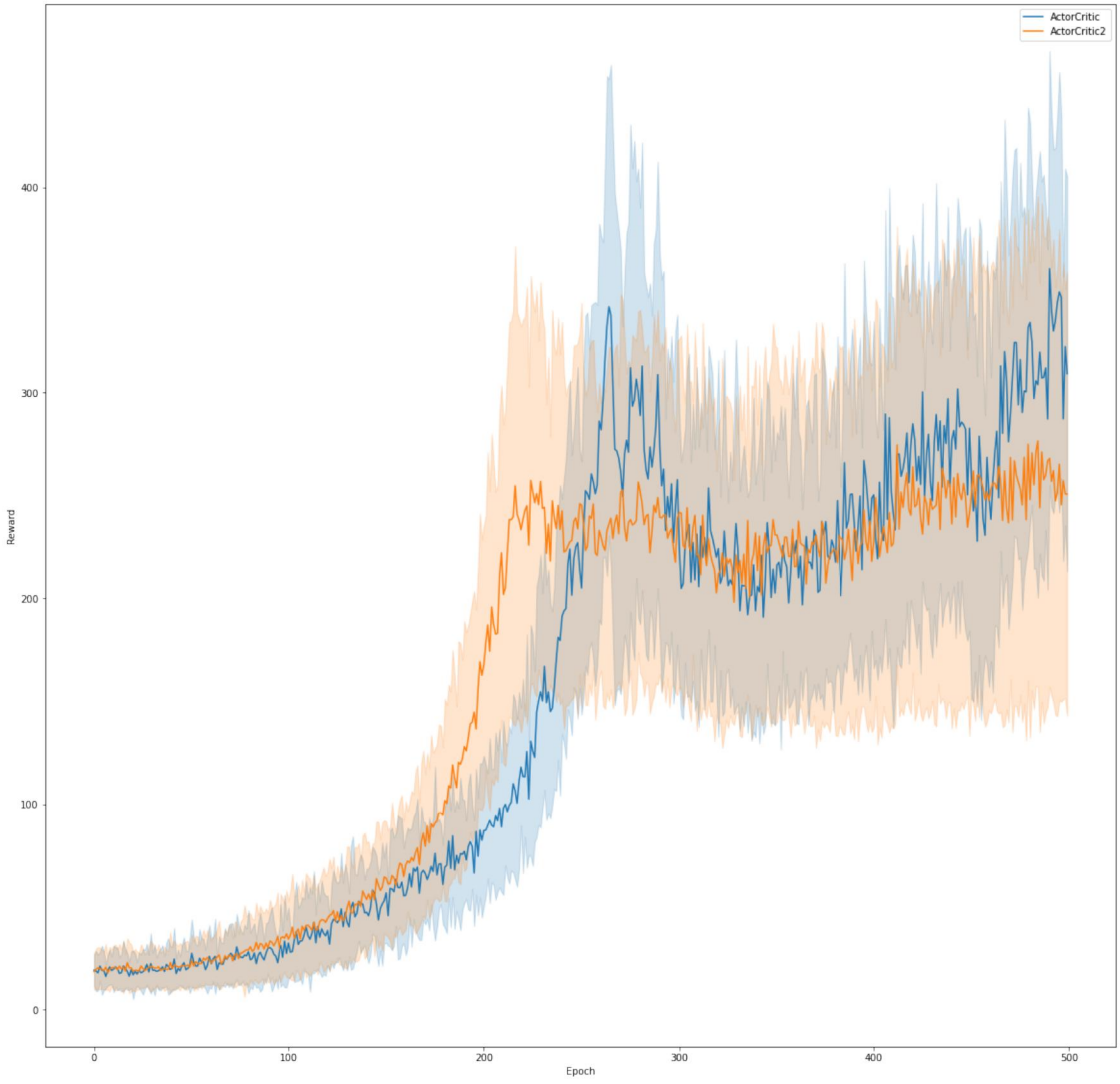
```
Iteration 448/500: rewards 249.06 +/- 98.3
Iteration 449/500: rewards 234.51 +/- 81.65
Iteration 450/500: rewards 241.23 +/- 107.48
Iteration 451/500: rewards 249.94 +/- 95.88
Iteration 452/500: rewards 261.85 +/- 113.38
Iteration 453/500: rewards 244.38 +/- 103.43
Iteration 454/500: rewards 259.67 +/- 102.59
Iteration 455/500: rewards 259.77 +/- 117.1
Iteration 456/500: rewards 257.01 +/- 101.24
Iteration 457/500: rewards 253.41 +/- 108.97
Iteration 458/500: rewards 247.86 +/- 109.62
Iteration 459/500: rewards 251.46 +/- 110.85
Iteration 460/500: rewards 247.54 +/- 101.35
Iteration 461/500: rewards 252.94 +/- 108.35
Iteration 462/500: rewards 256.05 +/- 107.55
Iteration 463/500: rewards 255.77 +/- 103.81
Iteration 464/500: rewards 252.91 +/- 106.82
Iteration 465/500: rewards 264.18 +/- 119.6
Iteration 466/500: rewards 256.3 +/- 110.77
Iteration 467/500: rewards 237.81 +/- 99.94
Iteration 468/500: rewards 261.85 +/- 122.92
Iteration 469/500: rewards 243.91 +/- 97.32
Iteration 470/500: rewards 236.65 +/- 96.45
Iteration 471/500: rewards 268.34 +/- 107.25
Iteration 472/500: rewards 237.98 +/- 102.64
Iteration 473/500: rewards 266.73 +/- 118.19
Iteration 474/500: rewards 259.67 +/- 111.59
Iteration 475/500: rewards 256.34 +/- 113.4
Iteration 476/500: rewards 253.12 +/- 99.5
Iteration 477/500: rewards 245.27 +/- 104.75
Iteration 478/500: rewards 268.47 +/- 118.95
Iteration 479/500: rewards 238.13 +/- 106.54
Iteration 480/500: rewards 274.72 +/- 114.07
Iteration 481/500: rewards 247.95 +/- 115.0
Iteration 482/500: rewards 270.6 +/- 114.53
Iteration 483/500: rewards 255.08 +/- 108.41
Iteration 484/500: rewards 270.84 +/- 112.8
Iteration 485/500: rewards 276.35 +/- 118.73
Iteration 486/500: rewards 243.95 +/- 108.36
Iteration 487/500: rewards 271.03 +/- 120.63
Iteration 488/500: rewards 257.53 +/- 117.63
Iteration 489/500: rewards 259.7 +/- 114.68
Iteration 490/500: rewards 266.6 +/- 118.64
Iteration 491/500: rewards 267.91 +/- 110.22
Iteration 492/500: rewards 256.73 +/- 103.69
Iteration 493/500: rewards 261.98 +/- 112.28
Iteration 494/500: rewards 247.29 +/- 103.33
Iteration 495/500: rewards 251.27 +/- 107.98
```

```
Iteration 496/500: rewards 265.07 +/- 113.97
Iteration 497/500: rewards 245.35 +/- 95.06
Iteration 498/500: rewards 257.05 +/- 105.62
Iteration 499/500: rewards 250.69 +/- 98.7
Iteration 500/500: rewards 250.61 +/- 107.1
```



[ ]:
```
# You will be graded on the output of this cell; So kindly run it
fig, ax = plt.subplots()
BaseAgent.plot_rewards(ActorCritic_sum_rewards , ax)
BaseAgent.plot_rewards(ActorCritic_rewards2, ax)
plt.rcParams['figure.figsize'] = [20, 20]
plt.legend(labels=['ActorCritic', 'ActorCritic2'])
```

[ ]:
```
<matplotlib.legend.Legend at 0x7fec348367d0>
```



In my case, I have done a mistake in my code and the resulting AC is not better than the original AC (before hyperparameter tuning) Since we didn't have much time for the AC part and the run took too much time, I wasn't able to rerun the grid search to get the right results. I hope you will be understanding, Thanks

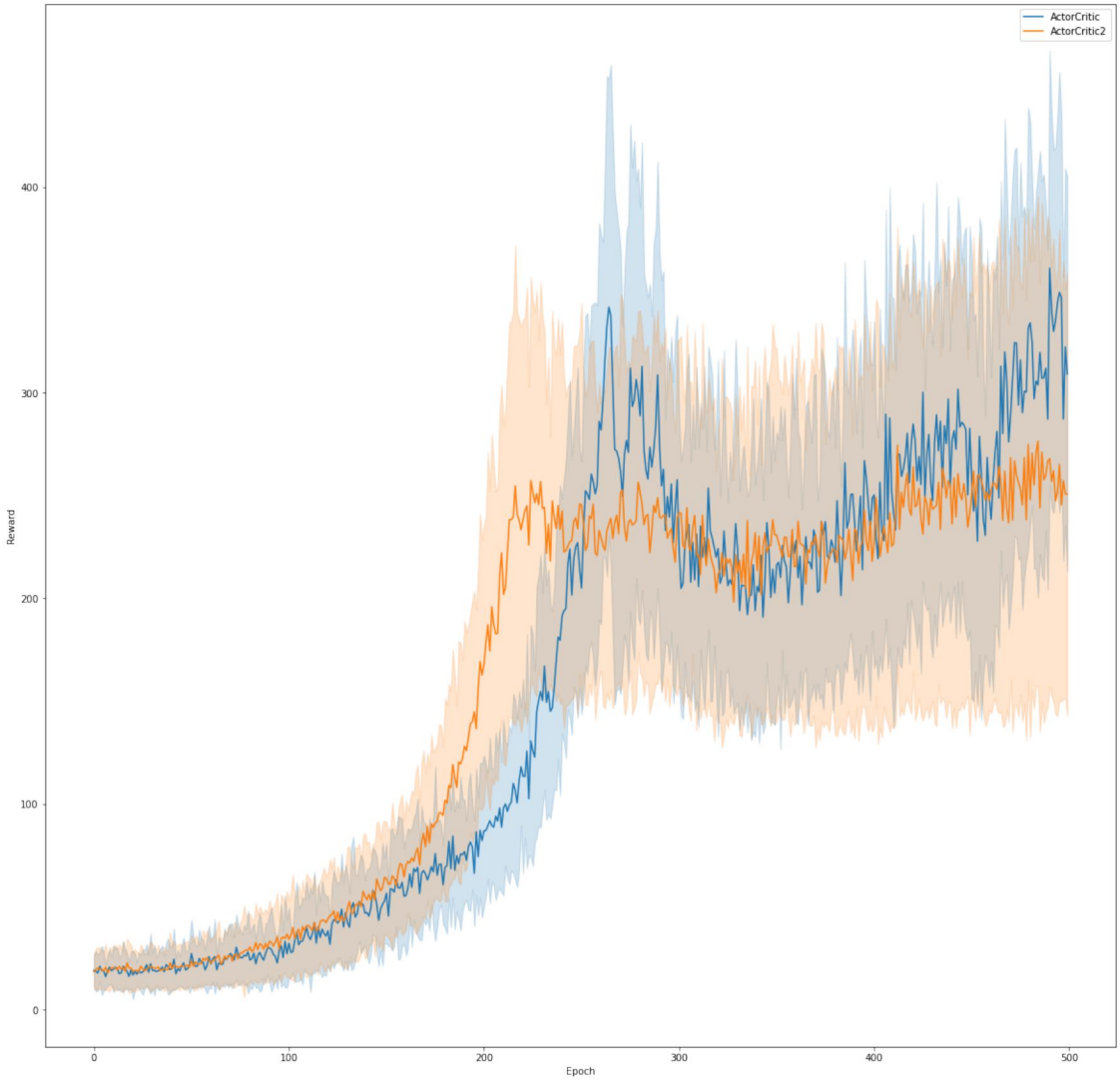
5.0.5 Qn 2.5: Compare and plot 'REINFORCEv2+B' method and 'ACTOR-CRITIC' method. [5 Marks]

Report your observations and provide explanations for the same.

[ ]:
```
# You will be graded on the output of this cell; So kindly run it
fig, ax = plt.subplots()
BaseAgent.plot_rewards(REINFORCEv2PlusBaselineAgent_rewards, ax)
BaseAgent.plot_rewards(ActorCritic_rewards2, ax)
plt.rcParams['figure.figsize'] = [20, 20]
```

4.4 Qn 2.4: Challenge! Can you tweak the hyperparameters of Actor-Critic to achieve better performance? Compare your results againts what you already have in section 3.1, in a single plot. **5 / 5**

✓ **- 0 pts** Correct



In my case, I have done a mistake in my code and the resulting AC is not better than the original AC (before hyperparameter tuning) Since we didn't have much time for the AC part and the run took too much time, I wasn't able to rerun the grid search to get the right results. I hope you will be understanding, Thanks

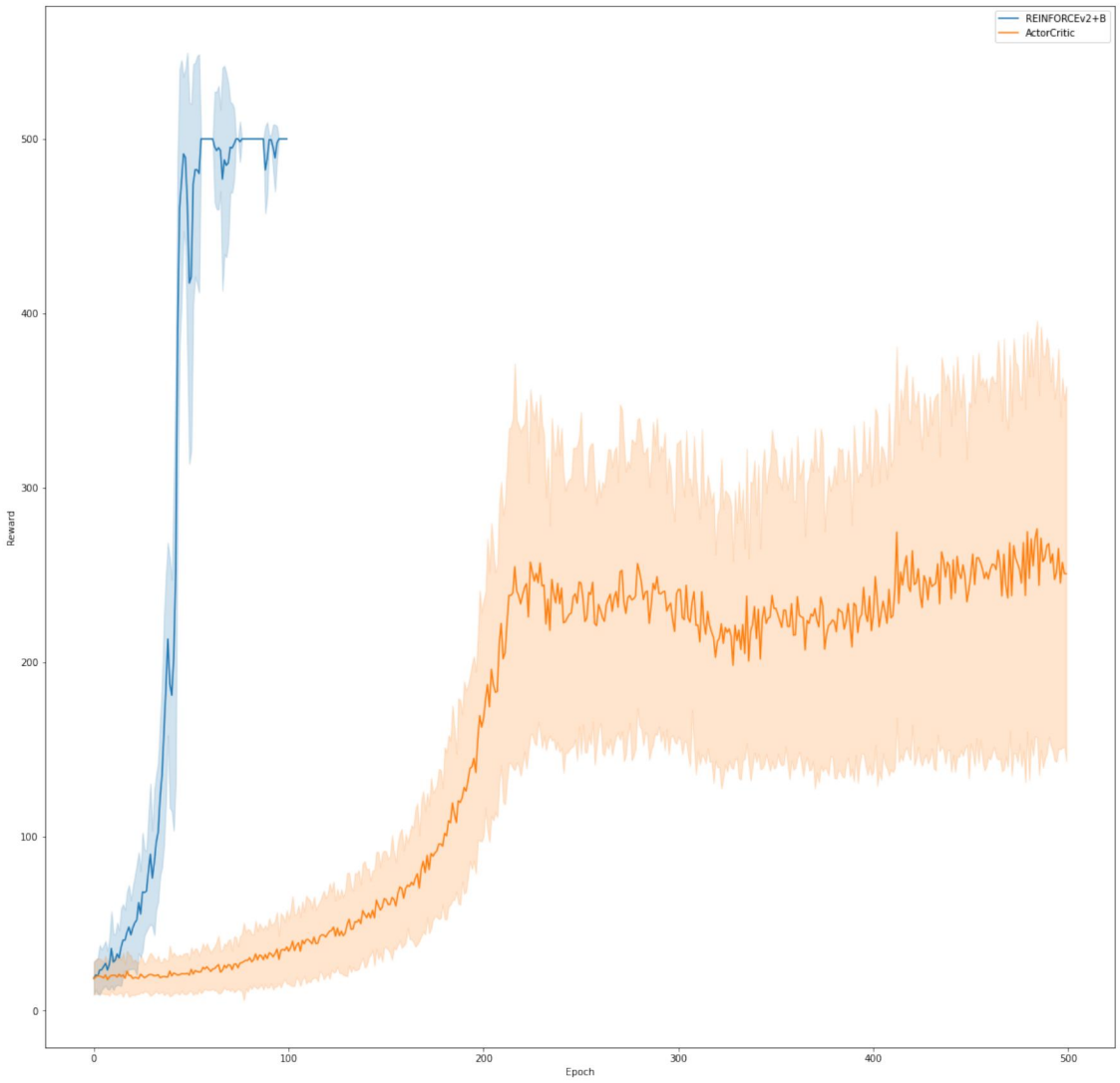
5.0.5 Qn 2.5: Compare and plot 'REINFORCEv2+B' method and 'ACTOR-CRITIC' method. [5 Marks]

Report your observations and provide explanations for the same.

[ ]:
```
# You will be graded on the output of this cell; So kindly run it
fig, ax = plt.subplots()
BaseAgent.plot_rewards(REINFORCEv2PlusBaselineAgent_rewards, ax)
BaseAgent.plot_rewards(ActorCritic_rewards2, ax)
plt.rcParams['figure.figsize'] = [20, 20]
```

```
plt.legend(labels=['REINFORCEv2+B', 'ActorCritic'])
```

```
[ ]: <matplotlib.legend.Legend at 0x7fec31cd1d10>
```



It is clear that in our case, Reinforce+Baseline is way better than the AC algorithm. Reinforce+B takes less than 60 epochs to converge toward 500 while Ac only converges to ~230 after 200 epochs. This shows that more complex algorithms aren't always better than simpler algorithm. Another observation that we can realize is that here AC has a bigger variance than Reinforce+B

4.5 Qn 2.5: Compare and plot `REINFORCEv2+B' method and 'ACTOR-CRITIC' method **4 / 5**

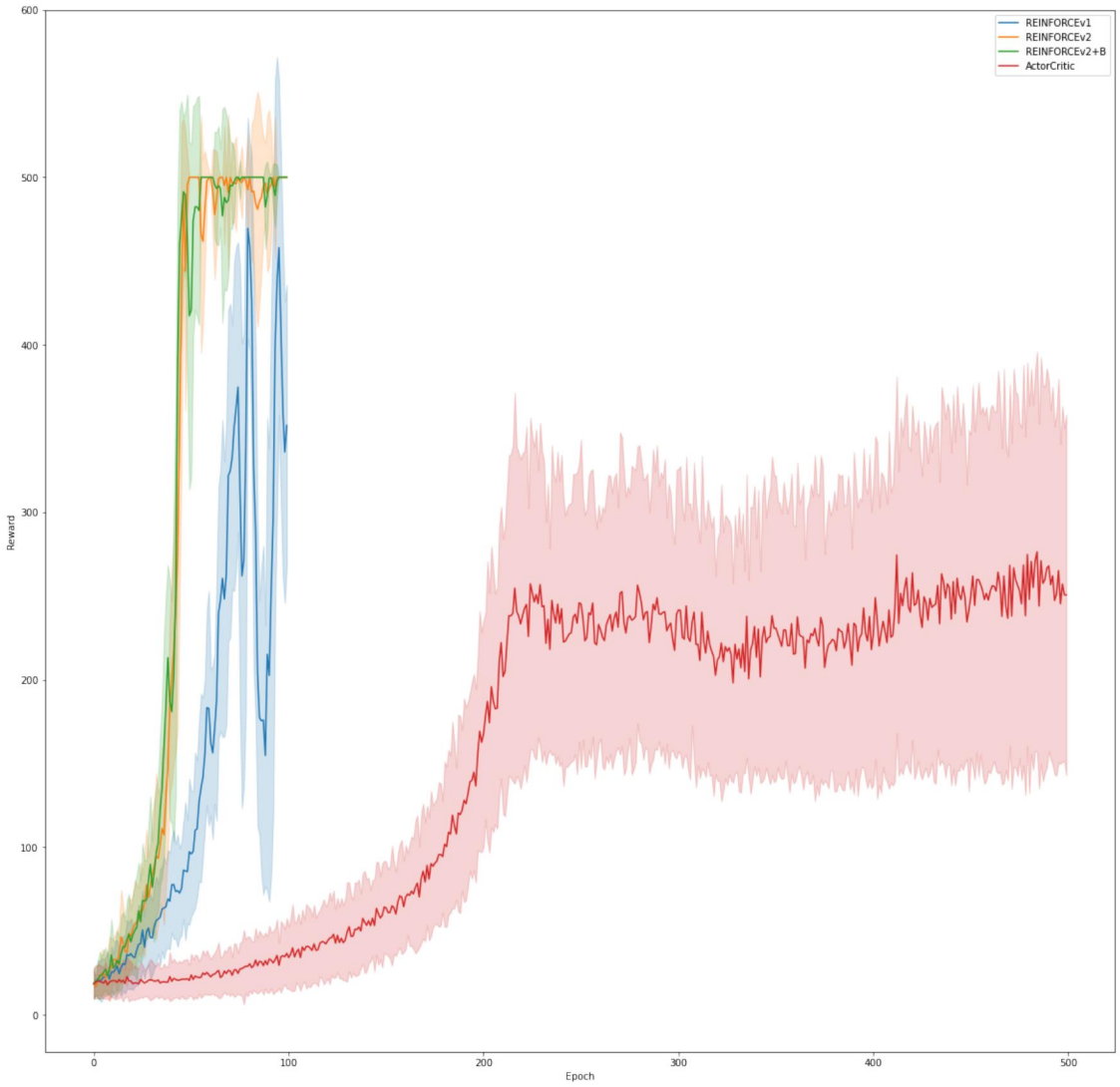
✓ **- 0 pts** Correct

✓ **- 1 pts** Poor AC performance

5.0.6 Qn 2.6: Plot all methods [2 Marks]

[ ]:
```
# You will be graded on the output of this cell; So kindly run it
fig, ax = plt.subplots()
BaseAgent.plot_rewards(REINFORCEv1_rewards, ax)
BaseAgent.plot_rewards(REINFORCEv2_rewards, ax)
BaseAgent.plot_rewards(REINFORCEv2PlusBaselineAgent_rewards, ax)
BaseAgent.plot_rewards(ActorCritic_rewards2, ax)
plt.rcParams['figure.figsize'] = [20, 20]
plt.legend(labels=['REINFORCEv1', 'REINFORCEv2', 'REINFORCEv2+B',␣
↪'ActorCritic'])
```

[ ]:
```
<matplotlib.legend.Legend at 0x7fec350c3fd0>
```



4.6 Qn 2.6: Plot all methods **2 / 2**

✓ **- 0 pts** **Correct**