INF8953DE: Assignment 3 - Policy Gradient Methods

Due on : Nov 19, 10:00 pm

Policy Gradient Methods

In this assignment, you will code RL agents that will learn a parameterized policy to solve the [CartPole](#) task. Simply put, the CartPole task is to balance a pole on a cart. Towards that, you will code and analyze different policy gradient methods that can maximize the objective of balancing the pole on cart without falling for maximum time.

Environment Details

For this assignment, we will use the [CartPole-v1](#) environment provided in the [OpenAI-Gym](#). The following is a description of the task as given in the gym:

CartPole Task: *A pole is attached by an un-actuated joint to a cart, which moves along a frictionless track. The pendulum starts upright, and the goal is to prevent it from falling over by increasing and reducing the cart's velocity.*

Detailed description and code can be found [here](#).

Policy Gradient Skeleton Code

Pseudocode

For this assignment you are asked to implement the policy gradient algorithms in the following pseudocode style.

Input: a differentiable policy network $\pi_\theta \in \mathcal{R}^d$ \ **Algorithm parameters:**

1. α: step size > 0
2. n_iterations: number of gradient updates > 0
3. n_episodes: number of episodes per gradient update > 0

Initialize policy parameters

loop for n*iterations: \   sample a dataset of episodes according to $\p\i\theta$* \

    # compute policy gradient \ $\nabla_\theta J(\theta) = \sum_j$ \   # where the first summation is over the $\sum_t \psi_{jt} \nabla_\theta ln \pi_\theta(a_t^j|s_t^j)$ episodes and the second summation is over the trajectory of the episode. \

    # update policy parameters \ $\theta_i = \theta_i + \alpha \nabla_\theta J(\theta)$

Please note that this version of the algorithm is different from the one you will see in the textbook. Specifically, instead of computing the gradient for each episode, we collect a batch of episodes and then compute our gradient using this entire batch (or dataset).

Installations and imports

```
In [1]:
```

```
# Run this cell

# install dependencies
!pip install torch torchvision pyvirtualdisplay matplotlib seaborn pandas numpy pathlib
gym
!sudo apt-get install xvfb
```

```
Requirement already satisfied: torch in /usr/local/lib/python3.7/dist-packages (1.10.0+cu
111)
Requirement already satisfied: torchvision in /usr/local/lib/python3.7/dist-packages (0.1
1.1+cu111)
Collecting pyvirtualdisplay
  Downloading PyVirtualDisplay-2.2-py3-none-any.whl (15 kB)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.7/dist-packages (3.2.
2)
Requirement already satisfied: seaborn in /usr/local/lib/python3.7/dist-packages (0.11.2)
Requirement already satisfied: pandas in /usr/local/lib/python3.7/dist-packages (1.1.5)
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (1.19.5)
Requirement already satisfied: pathlib in /usr/local/lib/python3.7/dist-packages (1.0.1)
Requirement already satisfied: gym in /usr/local/lib/python3.7/dist-packages (0.17.3)
Requirement already satisfied: typing-extensions in /usr/local/lib/python3.7/dist-package
s (from torch) (3.10.0.2)
Requirement already satisfied: pillow!=8.3.0,>=5.3.0 in /usr/local/lib/python3.7/dist-pac
kages (from torchvision) (7.1.2)
Collecting EasyProcess
  Downloading EasyProcess-0.3-py2.py3-none-any.whl (7.9 kB)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.7/dist-package
s (from matplotlib) (1.3.2)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.7/dist-packages (fr
om matplotlib) (0.11.0)
Requirement already satisfied: python-dateutil>=2.1 in /usr/local/lib/python3.7/dist-pack
ages (from matplotlib) (2.8.2)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in /usr/local/lib
/python3.7/dist-packages (from matplotlib) (3.0.6)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/dist-packages (from p
ython-dateutil>=2.1->matplotlib) (1.15.0)
Requirement already satisfied: scipy>=1.0 in /usr/local/lib/python3.7/dist-packages (from
seaborn) (1.4.1)
Requirement already satisfied: pytz>=2017.2 in /usr/local/lib/python3.7/dist-packages (fr
om pandas) (2018.9)
Requirement already satisfied: pyglet<=1.5.0,>=1.4.0 in /usr/local/lib/python3.7/dist-pac
kages (from gym) (1.5.0)
Requirement already satisfied: cloudpickle<1.7.0,>=1.2.0 in /usr/local/lib/python3.7/dist
-packages (from gym) (1.3.0)
Requirement already satisfied: future in /usr/local/lib/python3.7/dist-packages (from pyg
let<=1.5.0,>=1.4.0->gym) (0.16.0)
Installing collected packages: EasyProcess, pyvirtualdisplay
Successfully installed EasyProcess-0.3 pyvirtualdisplay-2.2
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following NEW packages will be installed:
  xvfb
0 upgraded, 1 newly installed, 0 to remove and 37 not upgraded.
Need to get 784 kB of archives.
After this operation, 2,270 kB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu bionic-updates/universe amd64 xvfb amd64 2:1.19.6-
1ubuntu4.9 [784 kB]
Fetched 784 kB in 1s (1,077 kB/s)
debconf: unable to initialize frontend: Dialog
debconf: (No usable dialog-like program is installed, so the dialog based frontend cannot
be used. at /usr/share/perl5/Debconf/FrontEnd/Dialog.pm line 76, <> line 1.)
debconf: falling back to frontend: Readline
debconf: unable to initialize frontend: Readline
debconf: (This frontend requires a controlling tty.)
debconf: falling back to frontend: Teletype
dpkg-preconfigure: unable to re-open stdin:
Selecting previously unselected package xvfb.
(Reading database ... 155222 files and directories currently installed.)
Preparing to unpack .../xvfb_2%3a1.19.6-1ubuntu4.9_amd64.deb ...
Unpacking xvfb (2:1.19.6-1ubuntu4.9) ...
Setting up xvfb (2:1.19.6-1ubuntu4.9) ...
```

```
Processing triggers for man-db (2.8.3-2ubuntu0.1) ...
```

In [2]:

```python
# Run this cell

# type hinting
from typing import Sequence, Tuple, Dict, Any, Optional

import numpy as np

# torch stuff
import torch
import torch.nn as nn
import torch.nn.functional as F
from torch import optim

# env
import gym
from gym.wrappers import Monitor

# data manipulation, colab dispaly, and plotting
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from pyvirtualdisplay import Display
from IPython import display as ipythondisplay
from IPython.display import clear_output

# misc util
import random, glob, base64, itertools
from pathlib import Path
from pprint import pprint
```

A util function to visualize the environment in colab:

In [3]:

```python
# Run this cell
def show_video(directory):
    html = []
    for mp4 in Path(directory).glob("*.mp4"):
        video_b64 = base64.b64encode(mp4.read_bytes())
        html.append('''<video alt="{}" autoplay
                      loop controls style="height: 400px;">
                      <source src="data:video/mp4;base64,{}" type="video/mp4" />
                 </video>'''.format(mp4, video_b64.decode('ascii')))
    ipythondisplay.display(ipythondisplay.HTML(data="<br>".join(html)))

display = Display(visible=0, size=(1400, 900))
display.start();
```

Building the Policy Network

The following piece of code is a util to make dense nets with PyTorch.

In [4]:

```python
# Run this cell
class Model(nn.Module):
    def __init__(self, features: Sequence[int]):
        """Fully-connected Network

        Args:
            features: a list of ints like: [input_dim, 16, 16, output_dim]
        """
        super(Model, self).__init__()
        layers = []
        for i in range(len(features) - 1):
            layers.append(
                nn.Linear(
                    in_features=features[i],
                    out_features=features[i + 1],
                    )
            )
            if i != len(features) - 2:
                layers.append(nn.ReLU())

        self.net = nn.Sequential(*layers)

    def forward(self, input):
        return self.net(input)
```

Building the Base Agent Class

In [5]:

```
# Run this cell
class BaseAgent(object):
    """ The base agent class function.
    """

    def __init__(self, config: Dict[str, Any]):
        """
        args:
            config: configuration dictionary
        """
        self.config = config

        # assert len(config['policy_layers']) > 0 # this won't allow linear models

        # environment
        self.env = gym.make(config['env_id'])
        self.gamma = config['gamma']

        # set seed
        np.random.seed(seed=config['seed'])
        self.env.seed(config['seed'])
        torch.manual_seed(config['seed'])

        # build policy model
        _policy_logits_model = Model(
            [self.env.observation_space.shape[0]] +
            config['policy_layers'] + # note that these are only the intermediate layers
            [self.env.action_space.n],
        )
        # NOTE: by design, policy model should take *batches* of states as input.
        # self.policy_model spits out the probability of each action
        self.policy_model = nn.Sequential(
            _policy_logits_model, nn.Softmax(dim=1),
        )
        self.policy_optimizer = torch.optim.Adam(
            self.policy_model.parameters(),
            lr=config['policy_learning_rate'],
        )
        self.monitor_env = Monitor(self.env, "./gym-results", force=True, video_callable
=lambda episode: True)

        if config['use_baseline']:
            self.value_model = Model(
              [self.env.observation_space.shape[0]] +
              self.config['value_layers'] + [1],
            )
            self.value_optimizer = torch.optim.Adam(self.value_model.parameters(), lr=co
nfig['value_learning_rate'])

    def _make_returns(self, rewards: np.ndarray):
        """ Compute the cumulative discounted rewards at each time step

        args:
            rewards: an array of step rewards

        returns:
            returns: an array of discounted returns from that timestep onward
        """
        returns = np.zeros_like(rewards)
        returns[-1] = rewards[-1]
        for t in reversed(range(len(rewards) - 1)):
            returns[t] = rewards[t] + self.gamma * returns[t + 1]
        return returns

    # Method to implement
    def optimize_model(self, n_episodes: int) -> np.ndarray:
        """ Takes a gradient step on policy (and value) parameters using
            `n_episodes` number of episodes. You'll need to implement
            this method for each part of this problem: namely, gather a
            dataset of size `n_episodes`, approximate the gradient using
            REINFORCE, and apply it to the model parameters.

        args:
            n_episodes: number of trajectories in dataset

        returns:
            returns: the total discounted reward of each trajectory/episode.
        """

        raise NotImplementedError

    def train(self, n_episodes: int, n_iterations: int, plot: bool = True) -> Sequence[np.ndarray]:
        """ Train.
        args:
            n_episodes: number of episodes for each gradient step
            n_iterations: determine training duration
        """

        rewards = []
        for it in range(n_iterations):
            rewards.append(self.optimize_model(n_episodes))
            print(f'Iteration {it + 1}/{n_iterations}: rewards {round(rewards[-1].mean(), 2)} +/- {round(rewards[-1].std(), 2)}')

        if plot:
            self.plot_rewards(rewards)

        return(rewards)

    @staticmethod
    def plot_rewards(rewards: Sequence[np.ndarray], ax: Optional[Any] = None):
        # Plotting
        r = pd.DataFrame((itertools.chain(*(itertools.product([i], rewards[i]) for i in range(len(rewards))))), columns=['Epoch', 'Reward'])
        if ax is None:
            sns.lineplot(x="Epoch", y="Reward", data=r, ci='sd');
        else:
            sns.lineplot(x="Epoch", y="Reward", data=r, ci='sd', ax=ax);

    def evaluate(self):
        """ Evaluate and visualize a single episode.
        """

        observation = self.monitor_env.reset()
        observation = torch.tensor(observation, dtype=torch.float)[None, :]
        reward_episode = 0
        done = False

        while not done:
            probs = self.policy_model.forward(observation)
            action = torch.multinomial(probs, 1)[0] # draw samples from dist
            observation, reward, done, info = self.monitor_env.step(int(action))
            observation = torch.tensor(observation, dtype=torch.float)[None, :]
            reward_episode += reward

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

Qn 1. REINFORCE ALGORITHM [65 Marks]

Qn 1.1 REINFORCE with episodal returns [25 Marks]

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.

In [6]:

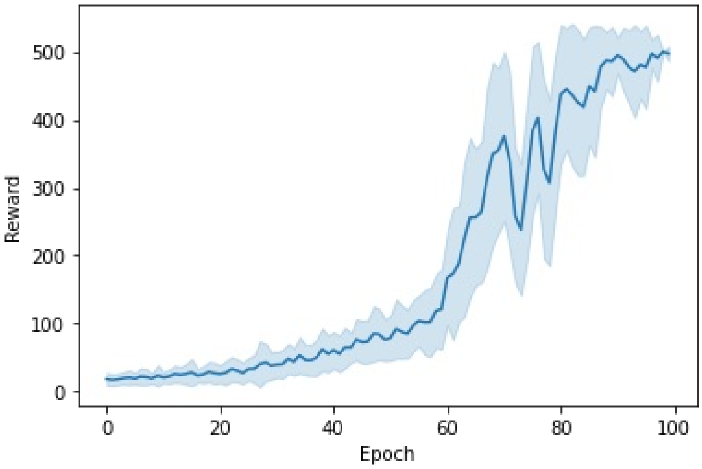
```
# Insert your code and run this cell
class REINFORCEv1Agent(BaseAgent):
    """ REINFORCE agent with total trajectory reward.
    """

    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 !
        total_rewards = np.zeros((n_episodes,))
        loss = 0.0
        for i in range(n_episodes):

            current_state = self.env.reset()
            done = False
            episodes = []
            while not done:

                action_probability = self.policy_model.forward(torch.FloatTensor(current
_state).unsqueeze(0))
                action = np.random.choice(np.array([0,1]), p = action_probability.data.n
umpy()[0])
                prev_state = current_state

                current_state, reward, done, extra = self.env.step(action)
                episodes.append((prev_state, action, reward))


            reward_batch = np.array([r for (s,a,r) in episodes])

            expected_return = self._make_returns(reward_batch)
            total_rewards[i] = sum(reward_batch)

            expected_return_batch = torch.FloatTensor(expected_return)
            expected_returns = expected_return[0] * torch.ones_like(expected_return_batc
h)
            state_batch = torch.Tensor([s for (s,a,r) in episodes])
            action_batch = torch.Tensor([a for (s,a,r) in episodes])

            prediction_batch = self.policy_model.forward(state_batch)
            action_selected_batch = prediction_batch.gather(dim = 1, index = action_batc
h.long().view(-1,1)).squeeze()


            loss -= torch.sum(torch.log(action_selected_batch) * expected_return[0])

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

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

In [7]:

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

# 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)
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:56: UserWarning: Creating a
tensor from a list of numpy.ndarrays is extremely slow. Please consider converting the li
st to a single numpy.ndarray with numpy.array() before converting to a tensor. (Triggered
internally at  ../torch/csrc/utils/tensor_new.cpp:201.)
```

```
Iteration 1/100: rewards 18.32 +/- 9.27
Iteration 2/100: rewards 16.84 +/- 8.34
Iteration 3/100: rewards 17.9 +/- 8.31
Iteration 4/100: rewards 19.64 +/- 9.31
Iteration 5/100: rewards 20.52 +/- 10.7
Iteration 6/100: rewards 19.04 +/- 8.63
Iteration 7/100: rewards 22.04 +/- 12.4
Iteration 8/100: rewards 21.42 +/- 11.87
Iteration 9/100: rewards 19.28 +/- 7.65
Iteration 10/100: rewards 23.44 +/- 15.43
Iteration 11/100: rewards 20.8 +/- 8.2
Iteration 12/100: rewards 22.18 +/- 10.32
Iteration 13/100: rewards 25.76 +/- 12.51
Iteration 14/100: rewards 24.36 +/- 11.7
Iteration 15/100: rewards 25.8 +/- 15.26
Iteration 16/100: rewards 28.2 +/- 19.89
Iteration 17/100: rewards 23.44 +/- 9.81
Iteration 18/100: rewards 24.86 +/- 12.29
Iteration 19/100: rewards 29.04 +/- 15.09
Iteration 20/100: rewards 26.92 +/- 15.64
Iteration 21/100: rewards 25.76 +/- 13.08
Iteration 22/100: rewards 27.46 +/- 17.69
Iteration 23/100: rewards 33.04 +/- 18.04
Iteration 24/100: rewards 30.78 +/- 17.04
Iteration 25/100: rewards 27.06 +/- 15.67
Iteration 26/100: rewards 32.94 +/- 16.56
Iteration 27/100: rewards 33.42 +/- 20.72
Iteration 28/100: rewards 40.28 +/- 33.44
Iteration 29/100: rewards 42.9 +/- 27.27
Iteration 30/100: rewards 38.1 +/- 20.43
Iteration 31/100: rewards 39.64 +/- 18.88
Iteration 32/100: rewards 40.14 +/- 20.63
Iteration 33/100: rewards 48.42 +/- 21.83
Iteration 34/100: rewards 43.64 +/- 19.6
Iteration 35/100: rewards 53.38 +/- 26.56
Iteration 36/100: rewards 46.32 +/- 21.94
Iteration 37/100: rewards 46.1 +/- 22.84
Iteration 38/100: rewards 50.1 +/- 26.56
Iteration 39/100: rewards 61.66 +/- 31.21
Iteration 40/100: rewards 55.44 +/- 27.01
Iteration 41/100: rewards 60.54 +/- 26.98
Iteration 42/100: rewards 55.52 +/- 26.26
Iteration 43/100: rewards 64.6 +/- 29.83
Iteration 44/100: rewards 64.5 +/- 21.59
Iteration 45/100: rewards 76.4 +/- 30.13
Iteration 46/100: rewards 72.74 +/- 31.12
Iteration 47/100: rewards 73.42 +/- 31.48
Iteration 48/100: rewards 84.96 +/- 40.17
Iteration 49/100: rewards 84.16 +/- 37.09
Iteration 50/100: rewards 76.14 +/- 30.75
Iteration 51/100: rewards 78.4 +/- 32.53
Iteration 52/100: rewards 92.24 +/- 42.98
Iteration 53/100: rewards 87.54 +/- 38.51
Iteration 54/100: rewards 85.06 +/- 35.56
Iteration 55/100: rewards 97.28 +/- 37.63
Iteration 56/100: rewards 103.78 +/- 37.19
Iteration 57/100: rewards 101.84 +/- 47.6
Iteration 58/100: rewards 102.04 +/- 49.83
Iteration 59/100: rewards 118.8 +/- 53.69
Iteration 60/100: rewards 121.04 +/- 58.45
Iteration 61/100: rewards 167.6 +/- 68.15
Iteration 62/100: rewards 172.94 +/- 96.22
Iteration 63/100: rewards 188.04 +/- 84.63
Iteration 64/100: rewards 224.02 +/- 112.71
Iteration 65/100: rewards 256.26 +/- 116.1
Iteration 66/100: rewards 256.58 +/- 100.18
Iteration 67/100: rewards 264.24 +/- 102.65
Iteration 68/100: rewards 313.6 +/- 131.63
Iteration 69/100: rewards 349.48 +/- 133.48
Iteration 70/100: rewards 354.32 +/- 120.75
Iteration 71/100: rewards 376.06 +/- 123.43
Iteration 72/100: rewards 339.34 +/- 130.01
Iteration 73/100: rewards 257.84 +/- 98.12
Iteration 74/100: rewards 237.78 +/- 95.87
Iteration 75/100: rewards 308.22 +/- 114.71
Iteration 76/100: rewards 383.66 +/- 122.92
Iteration 77/100: rewards 403.18 +/- 110.29
Iteration 78/100: rewards 327.3 +/- 129.94
Iteration 79/100: rewards 306.98 +/- 120.76
Iteration 80/100: rewards 381.18 +/- 114.81
Iteration 81/100: rewards 437.56 +/- 101.05
Iteration 82/100: rewards 445.08 +/- 88.76
Iteration 83/100: rewards 436.52 +/- 104.14
Iteration 84/100: rewards 425.0 +/- 105.93
Iteration 85/100: rewards 418.62 +/- 99.37
Iteration 86/100: rewards 449.18 +/- 85.76
Iteration 87/100: rewards 441.66 +/- 95.98
Iteration 88/100: rewards 477.78 +/- 58.48
Iteration 89/100: rewards 487.44 +/- 41.8
Iteration 90/100: rewards 485.88 +/- 50.19
Iteration 91/100: rewards 494.9 +/- 26.39
Iteration 92/100: rewards 488.94 +/- 46.28
Iteration 93/100: rewards 477.1 +/- 54.17
Iteration 94/100: rewards 471.1 +/- 67.78
Iteration 95/100: rewards 480.94 +/- 48.8
Iteration 96/100: rewards 477.26 +/- 60.87
Iteration 97/100: rewards 497.1 +/- 20.3
Iteration 98/100: rewards 490.94 +/- 34.54
Iteration 99/100: rewards 500.0 +/- 0.0
Iteration 100/100: rewards 497.28 +/- 10.86
```



In [ ]:

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

```
Reward: 500.0
```

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

In REINFORCE, we assume a finite horizon, so the episodes end after some finite st

```
eps; because of that, we do not necessarily need a discount factor.
Also, we have in Sutton Book: "If there is discounting
(gamma < 1) it should be treated as a form of termination.
Additionally, Reward discount reduces variance while reducing the impact of distant
actions."

So here in this game, we have a termination so we can consider
that the future reward is as good as the immediate reward.
Furthermore, if we want to decrease variance in the Q-value
estimation, we can use discount factor < 1 ( like in
critic-actor algorithm)`
```

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]

```
Because at the early stage, the agent does not know how to play and it loose really
fast, but when it trained properly,
which happens at the end, it can play longer ( it can play in more steps so it coll
ects more data and it take longer to
compute gradient and update the nueral networks)
and get more rewards, so it takes longer to train too.
```

Qn 1.2 REINFORCE with returns [15 Marks]

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.

In [30]:

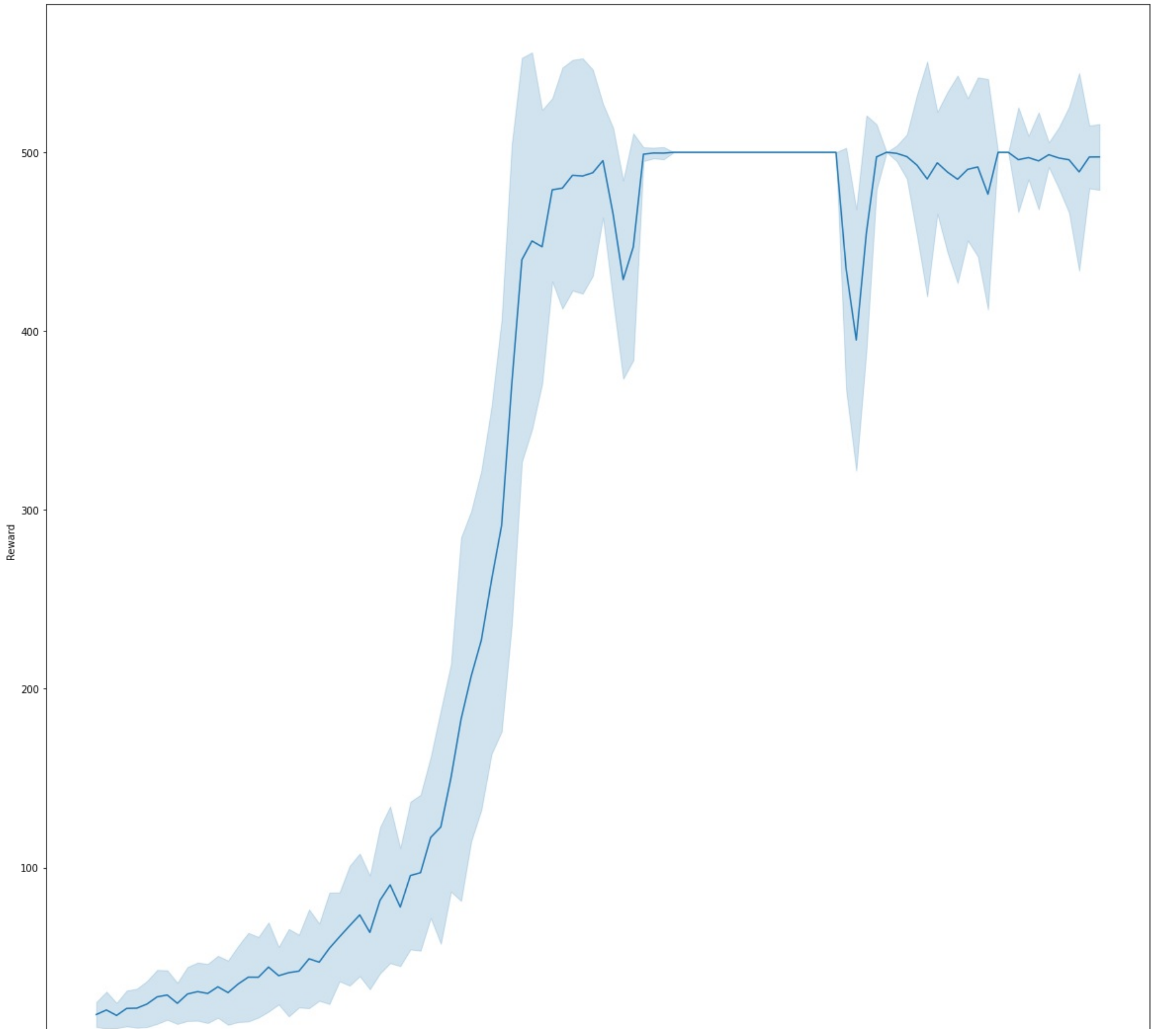
```
# Insert your code and run this cell
class REINFORCEv2Agent(BaseAgent):
    """ Not Vanilla REINFORCE Agent:
        *Not* vanilla, in the sense that we are now going to weight the action
        logprobs, proportionate to the onward return as opposed to the total
        episodic return.
    """

    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.
        """
        # ======================================================================

          # YOUR CODE HERE !
        total_rewards = np.zeros((n_episodes))
        loss = 0.0
        for i in range(n_episodes):

            current_state = self.env.reset()
            done = False
            episodes = []
            while not done:

                action_probability = self.policy_model.forward(torch.FloatTensor(current_state).unsqueeze(0))
                action = np.random.choice(np.array([0,1]), p = action_probability.data.numpy()[0])
                prev_state = current_state
                current_state, reward, done, extra = self.env.step(action)
                episodes.append((prev_state, action, reward))

            reward_batch = np.array([r for (s,a,r) in episodes])

            expected_return = self._make_returns( reward_batch)
            total_rewards[i] = sum(reward_batch)

            expected_returns_batch=torch.FloatTensor(expected_return)


            #expected_return_batch = torch.FloatTensor(expected_return)
            state_batch = torch.Tensor([s for (s,a,r) in episodes])
            action_batch = torch.Tensor([a for (s,a,r) in episodes])

            prediction_batch = self.policy_model.forward(state_batch)
            action_selected_batch = prediction_batch.gather(dim = 1, index = action_batch.long().view(-1,1)).squeeze()


            loss += - torch.sum(torch.log(action_selected_batch) * expected_returns_batch)




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

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

In [31]:

```
# 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)
```

```
Iteration 1/100: rewards 17.8 +/- 6.91
Iteration 2/100: rewards 20.32 +/- 10.14
Iteration 3/100: rewards 17.28 +/- 6.96
Iteration 4/100: rewards 21.2 +/- 9.9
Iteration 5/100: rewards 21.34 +/- 10.76
Iteration 6/100: rewards 23.64 +/- 12.72
Iteration 7/100: rewards 27.66 +/- 14.92
Iteration 8/100: rewards 28.7 +/- 13.69
Iteration 9/100: rewards 24.06 +/- 11.39
Iteration 10/100: rewards 29.26 +/- 14.91
Iteration 11/100: rewards 30.6 +/- 16.09
Iteration 12/100: rewards 29.58 +/- 16.35
Iteration 13/100: rewards 33.3 +/- 17.14
Iteration 14/100: rewards 30.04 +/- 17.77
Iteration 15/100: rewards 34.86 +/- 21.06
Iteration 16/100: rewards 38.68 +/- 24.51
Iteration 17/100: rewards 38.64 +/- 22.31
Iteration 18/100: rewards 44.34 +/- 24.67
Iteration 19/100: rewards 39.5 +/- 15.92
Iteration 20/100: rewards 41.2 +/- 24.2
Iteration 21/100: rewards 42.06 +/- 20.15
Iteration 22/100: rewards 48.96 +/- 27.33
Iteration 23/100: rewards 47.06 +/- 21.36
Iteration 24/100: rewards 54.82 +/- 30.83
Iteration 25/100: rewards 61.22 +/- 24.59
Iteration 26/100: rewards 67.44 +/- 33.13
Iteration 27/100: rewards 73.5 +/- 33.93
Iteration 28/100: rewards 63.7 +/- 31.44
Iteration 29/100: rewards 81.64 +/- 40.47
Iteration 30/100: rewards 90.32 +/- 43.35
Iteration 31/100: rewards 77.88 +/- 32.64
Iteration 32/100: rewards 95.52 +/- 40.88
Iteration 33/100: rewards 97.1 +/- 43.14
Iteration 34/100: rewards 116.8 +/- 44.56
Iteration 35/100: rewards 122.68 +/- 64.66
Iteration 36/100: rewards 150.1 +/- 62.93
Iteration 37/100: rewards 182.92 +/- 100.57
Iteration 38/100: rewards 207.02 +/- 91.39
Iteration 39/100: rewards 227.12 +/- 93.89
Iteration 40/100: rewards 260.5 +/- 96.12
Iteration 41/100: rewards 291.26 +/- 114.06
Iteration 42/100: rewards 370.06 +/- 132.91
Iteration 43/100: rewards 439.9 +/- 111.8
Iteration 44/100: rewards 450.4 +/- 104.35
Iteration 45/100: rewards 447.14 +/- 75.75
Iteration 46/100: rewards 479.02 +/- 50.69
Iteration 47/100: rewards 479.96 +/- 66.71
Iteration 48/100: rewards 487.14 +/- 63.87
Iteration 49/100: rewards 486.74 +/- 65.13
Iteration 50/100: rewards 488.56 +/- 57.08
Iteration 51/100: rewards 495.32 +/- 31.5
Iteration 52/100: rewards 465.5 +/- 47.74
Iteration 53/100: rewards 428.8 +/- 54.89
Iteration 54/100: rewards 447.1 +/- 62.79
Iteration 55/100: rewards 498.94 +/- 3.79
Iteration 56/100: rewards 499.58 +/- 2.94
Iteration 57/100: rewards 499.52 +/- 3.36
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 500.0 +/- 0.0
Iteration 64/100: rewards 500.0 +/- 0.0
Iteration 65/100: rewards 500.0 +/- 0.0
Iteration 66/100: rewards 500.0 +/- 0.0
Iteration 67/100: rewards 500.0 +/- 0.0
Iteration 68/100: rewards 500.0 +/- 0.0
Iteration 69/100: rewards 500.0 +/- 0.0
Iteration 70/100: rewards 500.0 +/- 0.0
Iteration 71/100: rewards 500.0 +/- 0.0
Iteration 72/100: rewards 500.0 +/- 0.0
Iteration 73/100: rewards 500.0 +/- 0.0
Iteration 74/100: rewards 500.0 +/- 0.0
Iteration 75/100: rewards 434.9 +/- 66.91
Iteration 76/100: rewards 395.02 +/- 72.38
Iteration 77/100: rewards 454.9 +/- 65.04
Iteration 78/100: rewards 497.42 +/- 18.06
Iteration 79/100: rewards 500.0 +/- 0.0
Iteration 80/100: rewards 499.38 +/- 4.34
Iteration 81/100: rewards 497.56 +/- 12.32
Iteration 82/100: rewards 492.78 +/- 39.1
Iteration 83/100: rewards 485.08 +/- 64.96
Iteration 84/100: rewards 494.12 +/- 28.19
Iteration 85/100: rewards 488.94 +/- 44.36
Iteration 86/100: rewards 484.92 +/- 57.42
Iteration 87/100: rewards 490.46 +/- 39.37
Iteration 88/100: rewards 491.8 +/- 49.55
Iteration 89/100: rewards 476.56 +/- 63.78
Iteration 90/100: rewards 500.0 +/- 0.0
Iteration 91/100: rewards 500.0 +/- 0.0
Iteration 92/100: rewards 495.86 +/- 28.98
Iteration 93/100: rewards 497.04 +/- 12.03
Iteration 94/100: rewards 495.18 +/- 26.81
Iteration 95/100: rewards 498.64 +/- 6.81
Iteration 96/100: rewards 496.8 +/- 16.98
Iteration 97/100: rewards 495.82 +/- 29.26
Iteration 98/100: rewards 489.02 +/- 54.69
Iteration 99/100: rewards 497.36 +/- 17.37
Iteration 100/100: rewards 497.4 +/- 18.2
```



In [ ]:

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

```
Reward: 500.0
```

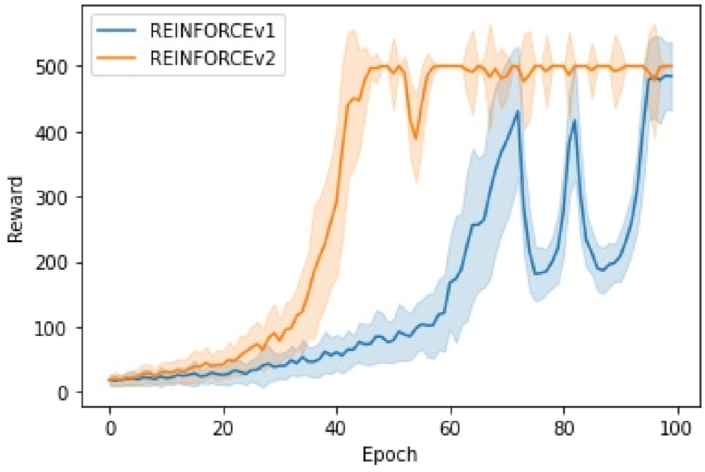
Qn 1.2.b: Plot and compare the performance of the `REINFORCEv1` and `REINFORCEv2` agents for $\gamma = 1$. Report your observations and provide explanations for the same. [5 Marks]

In [ ]:

```
# 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'])
```

Out[ ]:

```
<matplotlib.legend.Legend at 0x7fbd47aaa590>
```



```
REINFORCE is a monte carlo algorithm which is good for episodic case so it may have
high variance and produce slow learning.
REINFORCEv2 is converged faster and also is more stable and smoother, and less nois
y than REINFORCEv1. The difference between these two plots are the discounted retur
n we used.
In REINFORCEv1, we used expected returns that start from the first state for all st
eps in episodes; on the other hand, in REINFORCEv2, we used the expected return, wh
ich computed starting from the current state on the episodes. So in REINFORCEv1, in
stead of using discounted return for each step, we used the expected return of the
first step for all the steps on episodes, and it is clear that why it is noisier th
at second algorithm
```

Qn 1.3 REINFORCE WITH baseline 25 Marks]

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.

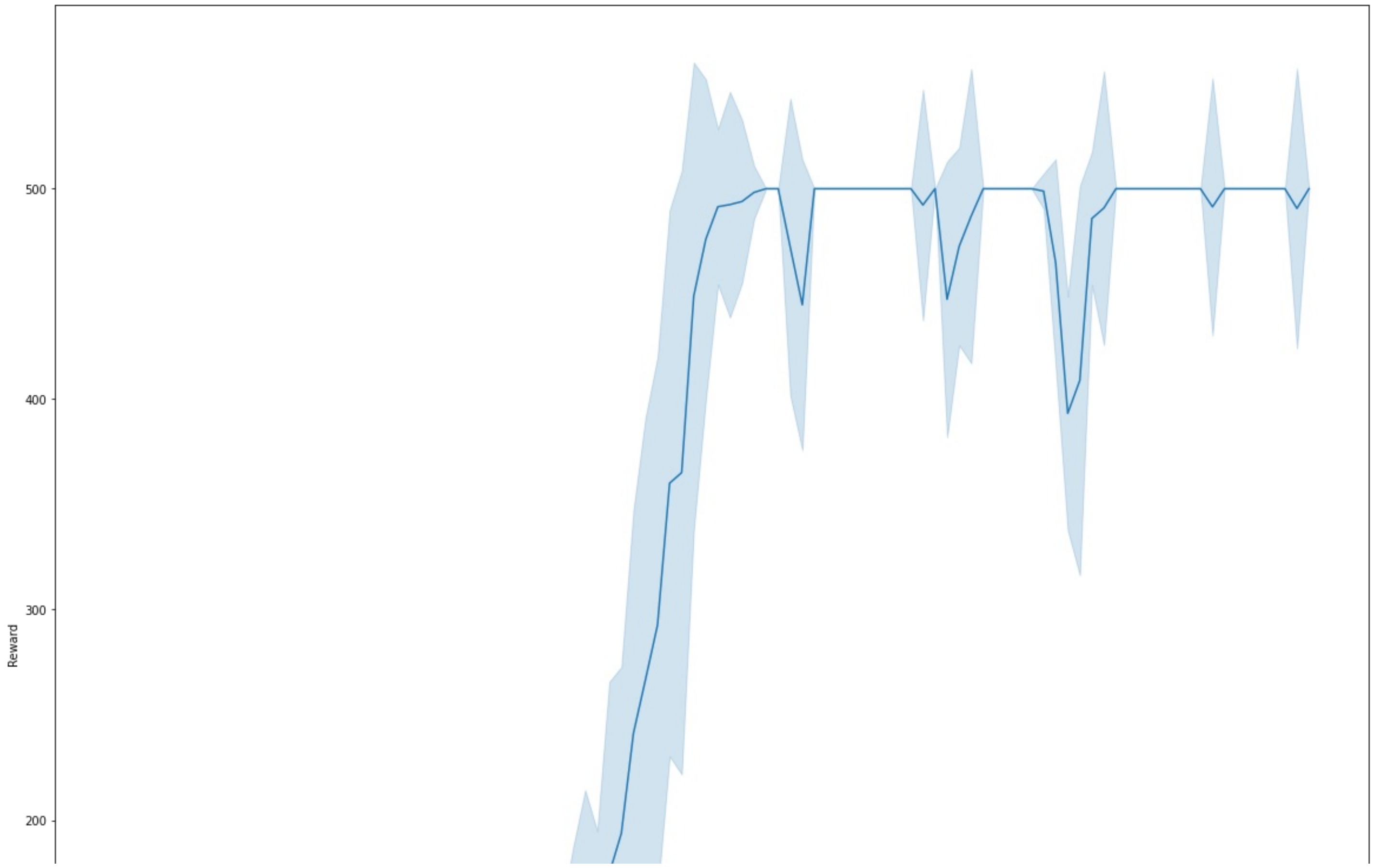
In [8]:

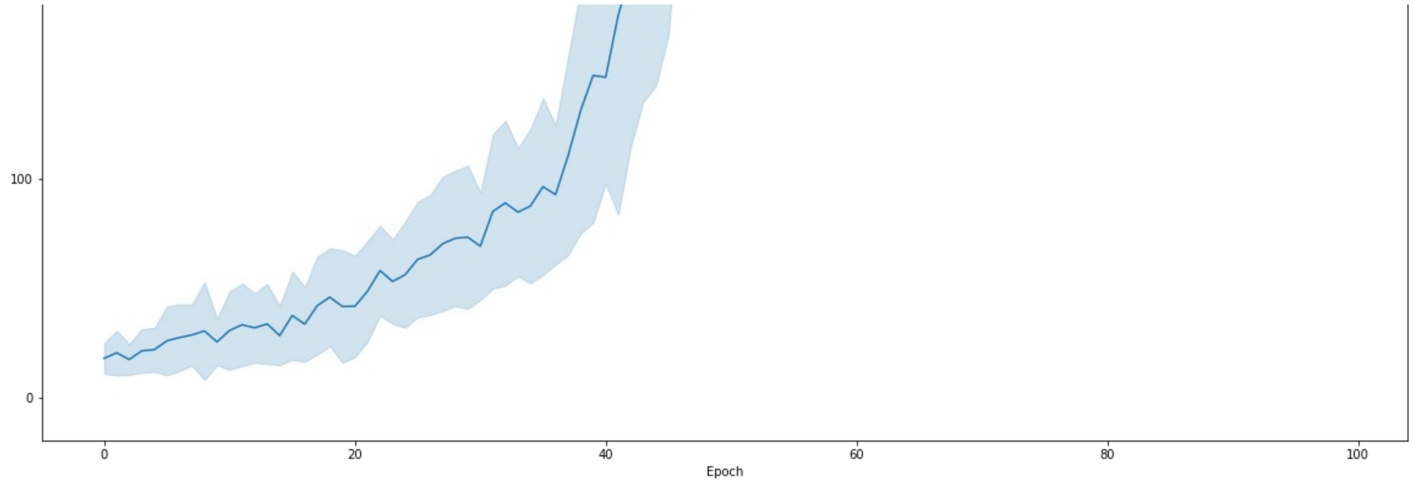
```python
# Insert your code and run this cell
class REINFORCEv2PlusBaselineAgent(BaseAgent):
    """ Baseline Agent:
        Here we try to reduce the variance by introducing a baseline, which is
        the value function in this case.
    """

    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.
        """
        # ====================================================================

          # YOUR CODE HERE !
        total_rewards = np.zeros((n_episodes))
        policy_loss = 0.0
        value_loss = 0.0
        for i in range(n_episodes):

            current_state = self.env.reset()
            done = False
            episodes = []
            while not done:
                #action_probability = self.policy_model(torch.from_numpy(current_state))
                #print(current_state)
                #print(torch.from_numpy(current_state).float())

                action_probability = self.policy_model.forward(torch.FloatTensor(current
_state).unsqueeze(0))

                action = np.random.choice(np.array([0,1]), p = action_probability.data.n
umpy()[0])
                prev_state = current_state
                current_state, reward, done, extra = self.env.step(action)
                episodes.append((prev_state, action, reward))
                #state = next_state

            reward_batch = np.array([r for (s,a,r) in episodes])



            expected_return = self._make_returns( reward_batch)
            total_rewards[i] = sum(reward_batch)

            expected_returns_batch=torch.FloatTensor(expected_return)


            #expected_return_batch = torch.FloatTensor(expected_return)
            state_batch = torch.Tensor([s for (s,a,r) in episodes])
            action_batch = torch.Tensor([a for (s,a,r) in episodes])

            value_baseline = self.value_model.forward(state_batch)


            prediction_batch = self.policy_model.forward(state_batch)
            action_selected_batch = prediction_batch.gather(dim = 1, index = action_batc
h.long().view(-1,1)).squeeze()


            policy_loss -= torch.sum(torch.log(action_selected_batch) * (expected_return
s_batch - value_baseline.detach()))
            value_loss += torch.sum(expected_returns_batch - value_baseline).pow(2)
        value_loss = value_loss / n_episodes




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

        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()
        return total_rewards
```

In [23]:

```
# 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)
```

```
Iteration 1/100: rewards 17.8 +/- 6.91
Iteration 2/100: rewards 20.32 +/- 10.14
Iteration 3/100: rewards 17.28 +/- 6.96
Iteration 4/100: rewards 21.2 +/- 9.9
Iteration 5/100: rewards 21.74 +/- 9.89
Iteration 6/100: rewards 25.78 +/- 15.53
Iteration 7/100: rewards 27.28 +/- 15.11
Iteration 8/100: rewards 28.48 +/- 13.84
Iteration 9/100: rewards 30.32 +/- 22.14
Iteration 10/100: rewards 25.3 +/- 10.51
Iteration 11/100: rewards 30.52 +/- 17.68
Iteration 12/100: rewards 33.14 +/- 18.68
Iteration 13/100: rewards 31.72 +/- 15.71
Iteration 14/100: rewards 33.52 +/- 18.08
Iteration 15/100: rewards 28.18 +/- 13.35
Iteration 16/100: rewards 37.36 +/- 20.03
Iteration 17/100: rewards 33.42 +/- 16.93
Iteration 18/100: rewards 41.9 +/- 22.16
Iteration 19/100: rewards 45.74 +/- 22.13
Iteration 20/100: rewards 41.48 +/- 25.45
Iteration 21/100: rewards 41.58 +/- 22.96
Iteration 22/100: rewards 48.42 +/- 22.76
Iteration 23/100: rewards 57.88 +/- 20.29
Iteration 24/100: rewards 52.94 +/- 19.06
Iteration 25/100: rewards 55.98 +/- 23.92
Iteration 26/100: rewards 62.98 +/- 26.14
Iteration 27/100: rewards 65.02 +/- 27.19
Iteration 28/100: rewards 70.16 +/- 30.36
Iteration 29/100: rewards 72.62 +/- 30.62
Iteration 30/100: rewards 73.14 +/- 32.38
Iteration 31/100: rewards 69.04 +/- 24.29
Iteration 32/100: rewards 84.92 +/- 34.9
Iteration 33/100: rewards 88.76 +/- 37.31
Iteration 34/100: rewards 84.6 +/- 28.98
Iteration 35/100: rewards 87.38 +/- 34.98
Iteration 36/100: rewards 96.26 +/- 39.93
Iteration 37/100: rewards 92.58 +/- 31.63
Iteration 38/100: rewards 110.3 +/- 44.82
Iteration 39/100: rewards 130.9 +/- 55.59
Iteration 40/100: rewards 147.06 +/- 66.49
Iteration 41/100: rewards 146.2 +/- 48.05
Iteration 42/100: rewards 174.56 +/- 90.26
Iteration 43/100: rewards 193.7 +/- 78.28
Iteration 44/100: rewards 241.16 +/- 104.97
Iteration 45/100: rewards 266.72 +/- 123.32
Iteration 46/100: rewards 292.68 +/- 126.07
Iteration 47/100: rewards 360.06 +/- 128.51
Iteration 48/100: rewards 365.08 +/- 141.88
Iteration 49/100: rewards 448.94 +/- 109.9
Iteration 50/100: rewards 475.98 +/- 75.24
Iteration 51/100: rewards 491.42 +/- 36.62
Iteration 52/100: rewards 492.42 +/- 53.06
Iteration 53/100: rewards 493.9 +/- 38.39
Iteration 54/100: rewards 498.22 +/- 12.46
Iteration 55/100: rewards 500.0 +/- 0.0
Iteration 56/100: rewards 500.0 +/- 0.0
Iteration 57/100: rewards 472.22 +/- 69.86
Iteration 58/100: rewards 444.8 +/- 68.35
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 500.0 +/- 0.0
Iteration 64/100: rewards 500.0 +/- 0.0
Iteration 65/100: rewards 500.0 +/- 0.0
Iteration 66/100: rewards 500.0 +/- 0.0
Iteration 67/100: rewards 500.0 +/- 0.0
Iteration 68/100: rewards 492.24 +/- 54.32
Iteration 69/100: rewards 500.0 +/- 0.0
Iteration 70/100: rewards 447.42 +/- 64.83
Iteration 71/100: rewards 472.56 +/- 46.47
Iteration 72/100: rewards 487.02 +/- 69.25
Iteration 73/100: rewards 500.0 +/- 0.0
Iteration 74/100: rewards 500.0 +/- 0.0
Iteration 75/100: rewards 500.0 +/- 0.0
Iteration 76/100: rewards 500.0 +/- 0.0
Iteration 77/100: rewards 500.0 +/- 0.0
Iteration 78/100: rewards 498.82 +/- 8.26
Iteration 79/100: rewards 464.88 +/- 48.67
Iteration 80/100: rewards 393.24 +/- 54.97
Iteration 81/100: rewards 408.82 +/- 91.51
Iteration 82/100: rewards 485.82 +/- 31.31
Iteration 83/100: rewards 490.8 +/- 64.4
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 500.0 +/- 0.0
Iteration 90/100: rewards 500.0 +/- 0.0
Iteration 91/100: rewards 500.0 +/- 0.0
Iteration 92/100: rewards 491.36 +/- 60.48
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 490.6 +/- 65.8
Iteration 100/100: rewards 500.0 +/- 0.0
```





In [9]:

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

```
Reward: 500.0
```

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

```
yes, it can affect the optimization process(learning rate)
when we want to estimate the gradient, the high variance is one of the major proble
ms, leading to slow convergence and noisy results.
Using the baseline, we expect it to reduce the variance, and the expected return wo
uld remain unbiased.
Moreover, it can increase the stability and speed of policy learning with REINFORCE
```

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.

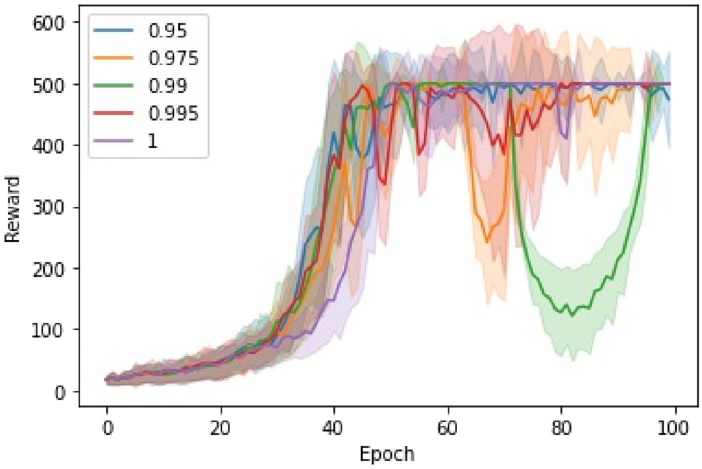
In [18]:

```
# Insert your code here and run this cell
# You will be graded on this output this cell, so kindly run it.
# keep the config
gammas = [0.95, 0.975, 0.99, 0.995, 1]
for i,gamma in enumerate(gammas):
    config = {
        'env_id': 'CartPole-v1',
        'seed': 8953,
        'gamma': gammas[i],
        '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)
plt.legend(labels=[0.95, 0.975, 0.99, 0.995, 1])
```

```
Iteration 1/100: rewards 17.8 +/- 6.91
Iteration 2/100: rewards 20.32 +/- 10.14
Iteration 3/100: rewards 17.28 +/- 6.96
Iteration 4/100: rewards 21.2 +/- 9.9
Iteration 5/100: rewards 20.94 +/- 10.87
Iteration 6/100: rewards 24.04 +/- 11.44
Iteration 7/100: rewards 27.66 +/- 14.77
Iteration 8/100: rewards 27.14 +/- 13.01
Iteration 9/100: rewards 26.5 +/- 12.02
Iteration 10/100: rewards 26.76 +/- 15.85
Iteration 11/100: rewards 32.1 +/- 20.61
Iteration 12/100: rewards 30.46 +/- 14.4
Iteration 13/100: rewards 28.28 +/- 12.56
Iteration 14/100: rewards 32.64 +/- 19.51
Iteration 15/100: rewards 34.26 +/- 22.86
Iteration 16/100: rewards 35.18 +/- 17.08
Iteration 17/100: rewards 34.24 +/- 17.11
Iteration 18/100: rewards 39.84 +/- 16.24
Iteration 19/100: rewards 41.22 +/- 23.37
Iteration 20/100: rewards 41.18 +/- 23.31
Iteration 21/100: rewards 48.6 +/- 24.18
Iteration 22/100: rewards 48.74 +/- 26.88
Iteration 23/100: rewards 51.34 +/- 22.4
Iteration 24/100: rewards 63.58 +/- 24.23
Iteration 25/100: rewards 67.74 +/- 37.1
Iteration 26/100: rewards 74.46 +/- 30.64
Iteration 27/100: rewards 72.24 +/- 43.76
Iteration 28/100: rewards 73.94 +/- 38.82
Iteration 29/100: rewards 82.7 +/- 42.99
Iteration 30/100: rewards 74.72 +/- 31.33
Iteration 31/100: rewards 89.4 +/- 30.09
Iteration 32/100: rewards 103.86 +/- 43.77
Iteration 33/100: rewards 124.12 +/- 59.61
Iteration 34/100: rewards 141.34 +/- 56.26
Iteration 35/100: rewards 181.84 +/- 89.96
Iteration 36/100: rewards 236.86 +/- 103.1
Iteration 37/100: rewards 252.78 +/- 108.37
Iteration 38/100: rewards 265.64 +/- 129.86
Iteration 39/100: rewards 260.24 +/- 129.39
Iteration 40/100: rewards 363.8 +/- 137.73
Iteration 41/100: rewards 419.56 +/- 118.09
Iteration 42/100: rewards 377.0 +/- 134.62
Iteration 43/100: rewards 463.76 +/- 85.88
Iteration 44/100: rewards 461.86 +/- 76.15
Iteration 45/100: rewards 394.68 +/- 117.62
Iteration 46/100: rewards 376.42 +/- 126.27
Iteration 47/100: rewards 386.28 +/- 116.73
Iteration 48/100: rewards 427.06 +/- 119.68
Iteration 49/100: rewards 476.36 +/- 66.64
Iteration 50/100: rewards 460.2 +/- 94.5
Iteration 51/100: rewards 465.56 +/-
Iteration 51/100: rewards 465.56 +/- 83.08
Iteration 52/100: rewards 464.0 +/- 85.44
Iteration 53/100: rewards 483.26 +/- 60.49
Iteration 54/100: rewards 497.68 +/- 16.24
Iteration 55/100: rewards 490.8 +/- 49.47
Iteration 56/100: rewards 498.4 +/- 11.2
Iteration 57/100: rewards 489.72 +/- 51.75
Iteration 58/100: rewards 479.6 +/- 51.82
Iteration 59/100: rewards 473.2 +/- 44.41
Iteration 60/100: rewards 481.7 +/- 43.44
Iteration 61/100: rewards 482.78 +/- 63.41
Iteration 62/100: rewards 491.74 +/- 57.82
Iteration 63/100: rewards 493.52 +/- 41.29
Iteration 64/100: rewards 481.7 +/- 89.75
Iteration 65/100: rewards 491.42 +/- 58.51
Iteration 66/100: rewards 491.48 +/- 59.64
Iteration 67/100: rewards 484.06 +/- 76.72
Iteration 68/100: rewards 500.0 +/- 0.0
Iteration 69/100: rewards 469.08 +/- 104.9
Iteration 70/100: rewards 500.0 +/- 0.0
Iteration 71/100: rewards 492.14 +/- 55.02
Iteration 72/100: rewards 498.96 +/- 7.28
Iteration 73/100: rewards 494.66 +/- 35.57
Iteration 74/100: rewards 482.36 +/- 86.55
Iteration 75/100: rewards 500.0 +/- 0.0
Iteration 76/100: rewards 490.68 +/- 65.24
Iteration 77/100: rewards 490.14 +/- 58.33
Iteration 78/100: rewards 500.0 +/- 0.0
Iteration 79/100: rewards 498.22 +/- 12.46
Iteration 80/100: rewards 491.68 +/- 58.24
Iteration 81/100: rewards 494.34 +/- 39.62
Iteration 82/100: rewards 492.44 +/- 52.92
Iteration 83/100: rewards 494.36 +/- 39.48
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 493.34 +/- 33.23
Iteration 88/100: rewards 489.14 +/- 53.6
Iteration 89/100: rewards 496.34 +/- 18.7
Iteration 90/100: rewards 500.0 +/- 0.0
Iteration 91/100: rewards 496.08 +/- 23.08
Iteration 92/100: rewards 498.76 +/- 8.68
Iteration 93/100: rewards 488.26 +/- 43.17
Iteration 94/100: rewards 500.0 +/- 0.0
Iteration 95/100: rewards 498.04 +/- 13.72
Iteration 96/100: rewards 495.14 +/- 25.42
Iteration 97/100: rewards 480.16 +/- 74.04
Iteration 98/100: rewards 489.86 +/- 56.08
Iteration 99/100: rewards 492.28 +/- 31.81
Iteration 100/100: rewards 473.6 +/- 78.11
Iteration 1/100: rewards 17.8 +/- 6.91
Iteration 2/100: rewards 20.32 +/- 10.14
Iteration 3/100: rewards 17.28 +/- 6.96
Iteration 4/100: rewards 21.2 +/- 9.9
Iteration 5/100: rewards 21.74 +/- 9.89
Iteration 6/100: rewards 23.24 +/- 12.01
Iteration 7/100: rewards 27.66 +/- 14.77
Iteration 8/100: rewards 24.36 +/- 11.43
Iteration 9/100: rewards 26.26 +/- 12.49
Iteration 10/100: rewards 30.74 +/- 16.9
Iteration 11/100: rewards 27.18 +/- 11.89
Iteration 12/100: rewards 29.46 +/- 15.68
Iteration 13/100: rewards 30.04 +/- 14.29
Iteration 14/100: rewards 31.98 +/- 19.67
Iteration 15/100: rewards 34.96 +/- 18.73
Iteration 16/100: rewards 37.32 +/- 20.06
Iteration 17/100: rewards 33.44 +/- 14.87
Iteration 18/100: rewards 42.54 +/- 22.37
Iteration 19/100: rewards 48.32 +/- 26.55
Iteration 20/100: rewards 39.76 +/- 20.41
Iteration 21/100: rewards 42.72 +/- 21.78
Iteration 22/100: rewards 44.96 +/- 26.57
Iteration 23/100: rewards 55.24 +/- 26.7
Iteration 24/100: rewards 52.48 +/- 19.35
Iteration 25/100: rewards 51.7 +/- 27.69
Iteration 26/100: rewards 73.04 +/- 35.5
Iteration 27/100: rewards 68.2 +/- 33.08
Iteration 28/100: rewards 66.02 +/- 28.39
Iteration 29/100: rewards 73.88 +/- 30.24
Iteration 30/100: rewards 84.5 +/- 45.05
Iteration 31/100: rewards 98.32 +/- 46.49
Iteration 32/100: rewards 87.4 +/- 36.6
Iteration 33/100: rewards 108.5 +/- 50.52
Iteration 34/100: rewards 121.92 +/- 59.87
Iteration 35/100: rewards 139.2 +/- 55.04
Iteration 36/100: rewards 154.62 +/- 71.81
Iteration 37/100: rewards 179.7 +/- 76.25
Iteration 38/100: rewards 192.3 +/- 65.8
Iteration 39/100: rewards 197.76 +/- 67.25
Iteration 40/100: rewards 227.88 +/- 88.4
Iteration 41/100: rewards 264.14 +/- 83.61
Iteration 42/100: rewards 318.24 +/- 124.06
Iteration 43/100: rewards 370.28 +/- 143.37
Iteration 44/100: rewards 277.84 +/- 123.21
Iteration 45/100: rewards 265.32 +/- 97.85
Iteration 46/100: rewards 375.22 +/- 114.3
Iteration 47/100: rewards 496.82 +/- 17.94
Iteration 48/100: rewards 478.78 +/- 61.99
Iteration 49/100: rewards 436.12 +/- 91.88
Iteration 50/100: rewards 413.36 +/- 100.2
Iteration 51/100: rewards 433.24 +/- 81.38
Iteration 52/100: rewards 458.18 +/- 83.39
Iteration 53/100: rewards 481.04 +/- 52.09
Iteration 54/100: rewards 496.28 +/- 26.04
Iteration 55/100: rewards 500.0 +/- 0.0
Iteration 56/100: rewards 497.64 +/- 16.52
Iteration 57/100: rewards 488.16 +/- 52.98
Iteration 58/100: rewards 499.4 +/- 3.79
Iteration 59/100: rewards 499.32 +/- 4.76
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 490.56 +/- 66.08
Iteration 64/100: rewards 478.5 +/- 91.36
Iteration 65/100: rewards 390.72 +/- 84.7
Iteration 66/100: rewards 289.34 +/- 71.13
Iteration 67/100: rewards 275.54 +/- 65.77
Iteration 68/100: rewards 240.78 +/- 99.98
Iteration 69/100: rewards 266.0 +/- 94.07
Iteration 70/100: rewards 271.64 +/- 116.28
Iteration 71/100: rewards 293.86 +/- 146.68
Iteration 72/100: rewards 427.94 +/- 101.6
Iteration 73/100: rewards 432.54 +/- 158.99
Iteration 74/100: rewards 466.14 +/- 106.73
Iteration 75/100: rewards 461.96 +/- 117.13
Iteration 76/100: rewards 454.06 +/- 131.39
Iteration 77/100: rewards 492.62 +/- 51.66
Iteration 78/100: rewards 463.9 +/- 122.63
Iteration 79/100: rewards 491.2 +/- 61.6
Iteration 80/100: rewards 490.5 +/- 66.5
Iteration 81/100: rewards 467.36 +/- 112.66
Iteration 82/100: rewards 464.04 +/- 122.02
Iteration 83/100: rewards 500.0 +/- 0.0
Iteration 84/100: rewards 470.48 +/- 110.75
Iteration 85/100: rewards 473.06 +/- 93.56
Iteration 86/100: rewards 479.8 +/- 68.4
Iteration 87/100: rewards 447.02 +/- 127.55
Iteration 88/100: rewards 474.76 +/- 94.29
Iteration 89/100: rewards 478.34 +/- 67.06
Iteration 90/100: rewards 466.76 +/- 93.52
Iteration 91/100: rewards 489.68 +/- 66.48
Iteration 92/100: rewards 490.46 +/- 66.78
Iteration 93/100: rewards 490.26 +/- 68.18
Iteration 94/100: rewards 500.0 +/- 0.0
Iteration 95/100: 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 492.3 +/- 53.9
Iteration 98/100: rewards 500.0 +/- 0.0
Iteration 99/100: rewards 500.0 +/- 0.0
Iteration 100/100: rewards 500.0 +/- 0.0
Iteration 1/100: rewards 17.8 +/- 6.91
Iteration 2/100: rewards 20.32 +/- 10.14
Iteration 3/100: rewards 17.28 +/- 6.96
Iteration 4/100: rewards 21.2 +/- 9.9
Iteration 5/100: rewards 21.74 +/- 9.89
Iteration 6/100: rewards 26.4 +/- 20.8
Iteration 7/100: rewards 24.64 +/- 12.38
Iteration 8/100: rewards 30.42 +/- 15.31
Iteration 9/100: rewards 28.38 +/- 14.49
Iteration 10/100: rewards 30.22 +/- 18.28
Iteration 11/100: rewards 26.9 +/- 11.78
Iteration 12/100: rewards 25.84 +/- 13.35
Iteration 13/100: rewards 31.48 +/- 18.77
Iteration 14/100: rewards 31.7 +/- 21.21
Iteration 15/100: rewards 38.36 +/- 22.72
Iteration 16/100: rewards 32.12 +/- 17.2
Iteration 17/100: rewards 40.16 +/- 23.42
Iteration 18/100: rewards 39.06 +/- 18.68
Iteration 19/100: rewards 40.26 +/- 18.73
Iteration 20/100: rewards 40.1 +/- 23.38
Iteration 21/100: rewards 48.52 +/- 25.91
Iteration 22/100: rewards 46.18 +/- 25.0
Iteration 23/100: rewards 49.72 +/- 23.9
Iteration 24/100: rewards 61.82 +/- 32.91
Iteration 25/100: rewards 62.68 +/- 31.66
Iteration 26/100: rewards 69.04 +/- 29.65
Iteration 27/100: rewards 66.98 +/- 27.41
Iteration 28/100: rewards 81.08 +/- 44.4
Iteration 29/100: rewards 81.52 +/- 35.77
Iteration 30/100: rewards 89.5 +/- 40.56
Iteration 31/100: rewards 111.64 +/- 65.86
Iteration 32/100: rewards 115.56 +/- 58.99
Iteration 33/100: rewards 130.48 +/- 73.96
Iteration 34/100: rewards 126.66 +/- 58.69
Iteration 35/100: rewards 140.88 +/- 65.93
Iteration 36/100: rewards 167.16 +/- 69.33
Iteration 37/100: rewards 195.66 +/- 99.31
Iteration 38/100: rewards 241.6 +/- 120.97
Iteration 39/100: rewards 273.1 +/- 150.16
Iteration 40/100: rewards 314.64 +/- 154.54
Iteration 41/100: rewards 355.16 +/- 130.87
Iteration 42/100: rewards 383.24 +/- 132.06
Iteration 43/100: rewards 430.18 +/- 93.75
Iteration 44/100: rewards 392.56 +/- 118.41
Iteration 45/100: rewards 459.84 +/- 107.58
Iteration 46/100: rewards 461.68 +/- 100.4
Iteration 47/100: rewards 458.38 +/- 89.11
Iteration 48/100: rewards 471.24 +/- 53.06
Iteration 49/100: rewards 454.32 +/- 98.7
Iteration 50/100: rewards 493.24 +/- 28.74
Iteration 51/100: rewards 499.98 +/- 0.14
Iteration 52/100: rewards 499.6 +/- 2.8
Iteration 53/100: rewards 482.98 +/- 31.88
Iteration 54/100: rewards 465.58 +/- 44.02
Iteration 55/100: rewards 428.08 +/- 57.99
Iteration 56/100: rewards 497.28 +/- 11.54
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 494.14 +/- 41.02
Iteration 61/100: rewards 500.0 +/- 0.0
Iteration 62/100: rewards 500.0 +/- 0.0
Iteration 63/100: rewards 500.0 +/- 0.0
Iteration 64/100: rewards 500.0 +/- 0.0
Iteration 65/100: rewards 498.9 +/- 7.7
Iteration 66/100: rewards 500.0 +/- 0.0
Iteration 67/100: rewards 496.64 +/- 23.52
Iteration 67/100: rewards 496.64 +/- 23.52
Iteration 68/100: rewards 500.0 +/- 0.0
Iteration 69/100: rewards 500.0 +/- 0.0
Iteration 70/100: rewards 500.0 +/- 0.0
Iteration 71/100: rewards 500.0 +/- 0.0
Iteration 72/100: rewards 499.74 +/- 1.31
Iteration 73/100: rewards 335.5 +/- 38.2
Iteration 74/100: rewards 259.78 +/- 19.88
Iteration 75/100: rewards 219.28 +/- 31.35
Iteration 76/100: rewards 189.06 +/- 48.86
Iteration 77/100: rewards 180.74 +/- 39.96
Iteration 78/100: rewards 156.8 +/- 58.77
Iteration 79/100: rewards 149.98 +/- 59.8
Iteration 80/100: rewards 132.38 +/- 69.45
Iteration 81/100: rewards 127.24 +/- 69.77
Iteration 82/100: rewards 139.44 +/- 62.92
Iteration 83/100: rewards 121.76 +/- 73.17
Iteration 84/100: rewards 134.68 +/- 65.16
Iteration 85/100: rewards 136.5 +/- 69.28
Iteration 86/100: rewards 133.8 +/- 73.0
Iteration 87/100: rewards 161.14 +/- 61.05
Iteration 88/100: rewards 165.1 +/- 63.91
Iteration 89/100: rewards 185.52 +/- 61.77
Iteration 90/100: rewards 181.36 +/- 72.86
Iteration 91/100: rewards 212.8 +/- 52.31
Iteration 92/100: rewards 223.62 +/- 67.61
Iteration 93/100: rewards 261.72 +/- 39.8
Iteration 94/100: rewards 296.22 +/- 44.59
Iteration 95/100: rewards 340.76 +/- 66.08
Iteration 96/100: rewards 422.58 +/- 76.37
Iteration 97/100: rewards 500.0 +/- 0.0
Iteration 98/100: rewards 494.1 +/- 41.3
Iteration 99/100: rewards 500.0 +/- 0.0
Iteration 100/100: rewards 500.0 +/- 0.0
Iteration 1/100: rewards 17.8 +/- 6.91
Iteration 2/100: rewards 20.32 +/- 10.14
Iteration 3/100: rewards 17.28 +/- 6.96
Iteration 4/100: rewards 21.2 +/- 9.9
Iteration 5/100: rewards 21.74 +/- 9.89
Iteration 6/100: rewards 26.08 +/- 16.35
Iteration 7/100: rewards 24.96 +/- 12.57
Iteration 8/100: rewards 30.42 +/- 15.65
Iteration 9/100: rewards 28.3 +/- 14.86
Iteration 10/100: rewards 27.0 +/- 17.64
Iteration 11/100: rewards 26.76 +/- 14.87
Iteration 12/100: rewards 33.26 +/- 18.39
Iteration 13/100: rewards 31.66 +/- 17.45
Iteration 14/100: rewards 32.48 +/- 15.4
Iteration 15/100: rewards 36.82 +/- 18.7
Iteration 16/100: rewards 36.8 +/- 20.91
Iteration 17/100: rewards 40.76 +/- 20.31
Iteration 18/100: rewards 44.22 +/- 26.66
Iteration 19/100: rewards 43.88 +/- 22.19
Iteration 20/100: rewards 44.9 +/- 23.16
Iteration 21/100: rewards 44.78 +/- 27.92
Iteration 22/100: rewards 52.32 +/- 27.36
Iteration 23/100: rewards 54.5 +/- 30.44
Iteration 24/100: rewards 55.62 +/- 26.08
Iteration 25/100: rewards 60.56 +/- 33.0
Iteration 26/100: rewards 61.3 +/- 27.2
Iteration 27/100: rewards 61.04 +/- 27.19
Iteration 28/100: rewards 68.76 +/- 26.9
Iteration 29/100: rewards 78.3 +/- 32.44
Iteration 30/100: rewards 83.5 +/- 30.7
Iteration 31/100: rewards 96.16 +/- 33.65
Iteration 32/100: rewards 121.46 +/- 53.67
Iteration 33/100: rewards 127.48 +/- 64.14
Iteration 34/100: rewards 143.54 +/- 54.86
Iteration 35/100: rewards 151.88 +/- 62.86
Iteration 36/100: rewards 193.44 +/- 91.57
Iteration 37/100: rewards 201.74 +/- 83.29
Iteration 38/100: rewards 212.12 +/- 91.6
Iteration 39/100: rewards 265.94 +/- 120.4
Iteration 40/100: rewards 350.02 +/- 136.49
Iteration 41/100: rewards 383.08 +/- 134.6
Iteration 42/100: rewards 361.36 +/- 125.33
Iteration 43/100: rewards 440.58 +/- 106.6
Iteration 44/100: rewards 471.04 +/- 83.3
Iteration 45/100: rewards 479.42 +/- 70.26
Iteration 46/100: rewards 496.46 +/- 20.81
Iteration 47/100: rewards 488.04 +/- 40.3
Iteration 48/100: rewards 459.9 +/- 74.53
Iteration 49/100: rewards 347.6 +/- 94.26
Iteration 50/100: rewards 335.2 +/- 98.54
Iteration 51/100: rewards 418.84 +/- 81.87
Iteration 52/100: rewards 493.02 +/- 26.2
Iteration 53/100: rewards 500.0 +/- 0.0
Iteration 54/100: rewards 500.0 +/- 0.0
Iteration 55/100: rewards 491.14 +/- 23.19
Iteration 56/100: rewards 383.88 +/- 66.5
Iteration 57/100: rewards 402.02 +/- 70.92
Iteration 58/100: rewards 493.0 +/- 45.17
Iteration 59/100: rewards 482.32 +/- 79.82
Iteration 60/100: rewards 486.7 +/- 70.26
Iteration 61/100: rewards 481.44 +/- 77.09
Iteration 62/100: rewards 476.1 +/- 82.17
Iteration 63/100: rewards 495.72 +/- 29.96
Iteration 64/100: rewards 491.48 +/- 59.64
Iteration 65/100: rewards 475.96 +/- 96.86
Iteration 66/100: rewards 468.48 +/- 107.72
Iteration 67/100: rewards 455.24 +/- 130.61
Iteration 68/100: rewards 435.06 +/- 145.99
Iteration 69/100: rewards 398.38 +/- 186.37
Iteration 70/100: rewards 416.38 +/- 178.63
Iteration 71/100: rewards 384.14 +/- 196.89
Iteration 72/100: rewards 473.94 +/- 104.02
Iteration 73/100: rewards 415.9 +/- 179.7
Iteration 74/100: rewards 415.54 +/- 177.19
Iteration 75/100: rewards 457.06 +/- 115.11
Iteration 76/100: rewards 424.64 +/- 161.36
Iteration 77/100: rewards 439.52 +/- 148.3
Iteration 78/100: rewards 466.8 +/- 93.18
Iteration 79/100: rewards 458.0 +/- 92.91
Iteration 80/100: rewards 471.06 +/- 73.33
Iteration 81/100: rewards 500.0 +/- 0.0
Iteration 82/100: rewards 492.18 +/- 54.74
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 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 491.26 +/- 61.18
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
Iteration 1/100: rewards 17.8 +/- 6.91
Iteration 2/100: rewards 20.32 +/- 10.14
Iteration 3/100: rewards 17.28 +/- 6.96
Iteration 4/100: rewards 21.2 +/- 9.9
Iteration 5/100: rewards 21.74 +/- 9.89
Iteration 6/100: rewards 25.78 +/- 15.53
Iteration 7/100: rewards 27.28 +/- 15.11
Iteration 8/100: rewards 28.48 +/- 13.84
Iteration 9/100: rewards 30.32 +/- 22.14
Iteration 10/100: rewards 25.3 +/- 10.51
Iteration 11/100: rewards 30.52 +/- 17.68
Iteration 12/100: rewards 33.14 +/- 18.68
Iteration 13/100: rewards 31.72 +/- 15.71
Iteration 14/100: rewards 33.52 +/- 18.08
Iteration 15/100: rewards 28.18 +/- 13.35
Iteration 16/100: rewards 37.36 +/- 20.03
Iteration 17/100: rewards 33.42 +/- 16.93
Iteration 18/100: rewards 41.9 +/- 22.16
Iteration 19/100: rewards 45.74 +/- 22.13
Iteration 20/100: rewards 41.48 +/- 25.45
Iteration 21/100: rewards 41.58 +/- 22.96
Iteration 22/100: rewards 48.42 +/- 22.76
Iteration 23/100: rewards 57.88 +/- 20.29
Iteration 24/100: rewards 52.94 +/- 19.06
Iteration 25/100: rewards 55.98 +/- 23.92
Iteration 26/100: rewards 62.98 +/- 26.14
Iteration 27/100: rewards 65.02 +/- 27.19
Iteration 28/100: rewards 70.16 +/- 30.36
Iteration 29/100: rewards 72.62 +/- 30.62
Iteration 30/100: rewards 73.14 +/- 32.38
Iteration 31/100: rewards 69.04 +/- 24.29
Iteration 32/100: rewards 84.92 +/- 34.9
Iteration 33/100: rewards 88.76 +/- 37.31
Iteration 34/100: rewards 84.6 +/- 28.98
Iteration 35/100: rewards 87.38 +/- 34.98
Iteration 36/100: rewards 96.26 +/- 39.93
Iteration 37/100: rewards 92.58 +/- 31.63
Iteration 38/100: rewards 110.3 +/- 44.82
Iteration 39/100: rewards 130.9 +/- 55.59
Iteration 40/100: rewards 147.06 +/- 66.49
Iteration 41/100: rewards 146.2 +/- 48.05
Iteration 42/100: rewards 174.56 +/- 90.26
Iteration 43/100: rewards 193.7 +/- 78.28
Iteration 44/100: rewards 241.16 +/- 104.97
Iteration 45/100: rewards 266.72 +/- 123.32
Iteration 46/100: rewards 292.68 +/- 126.07
Iteration 47/100: rewards 360.06 +/- 128.51
Iteration 48/100: rewards 365.08 +/- 141.88
Iteration 49/100: rewards 448.94 +/- 109.9
Iteration 50/100: rewards 475.26 +/- 70.5
Iteration 51/100: rewards 500.0 +/- 0.0
Iteration 52/100: rewards 495.74 +/- 22.61
Iteration 53/100: rewards 490.38 +/- 48.61
Iteration 54/100: rewards 483.88 +/- 69.79
Iteration 55/100: rewards 498.72 +/- 6.95
Iteration 56/100: rewards 475.6 +/- 70.67
Iteration 57/100: rewards 465.44 +/- 79.61
Iteration 58/100: rewards 462.94 +/- 82.73
Iteration 59/100: rewards 475.56 +/- 67.81
Iteration 60/100: rewards 487.32 +/- 54.13
Iteration 61/100: rewards 484.26 +/- 55.19
Iteration 62/100: rewards 488.14 +/- 36.96
Iteration 63/100: rewards 485.52 +/- 54.19
Iteration 64/100: rewards 488.4 +/- 46.77
Iteration 65/100: rewards 493.4 +/- 32.66
Iteration 66/100: rewards 497.66 +/- 16.38
Iteration 67/100: rewards 496.3 +/- 25.9
Iteration 68/100: rewards 500.0 +/- 0.0
Iteration 69/100: rewards 500.0 +/- 0.0
Iteration 70/100: rewards 493.9 +/- 42.7
Iteration 71/100: rewards 500.0 +/- 0.0
Iteration 72/100: rewards 500.0 +/- 0.0
Iteration 73/100: rewards 500.0 +/- 0.0
Iteration 74/100: rewards 500.0 +/- 0.0
Iteration 75/100: rewards 500.0 +/- 0.0
Iteration 76/100: rewards 499.66 +/- 2.38
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 419.7 +/- 63.33
Iteration 82/100: rewards 410.12 +/- 74.41
Iteration 83/100: rewards 494.48 +/- 18.66
Iteration 84/100: rewards 490.4 +/- 67.2
Iteration 85/100: rewards 500.0 +/- 0.0
Iteration 86/100: rewards 500.0 +/- 0.0
Iteration 87/100: rewards 498.16 +/- 12.88
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 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
```

Out[18]:

```
<matplotlib.legend.Legend at 0x7f3d35ed6f10>
```



The baseline function is a function which try to not change the expected value (so it would be unbiased), but it will decrease the variance.
Gamma = 1 starts to converge later than the other gammas, but it also converges really fast(around episode 60).
It's more stable and less noisy.(low biased) Also, it has high variance either.

Gamma = 0.995 starts to converge faster but is unstable(biased) and converges later than gamma = 1(around episode 80).
Compared to the other plot, it has less variance compare to Gamma =1.

gamma = 0.99. It's the most unstable(noisy-biased) plot and also converges really slow (around episode 100)

gamma = 0.975 it's converge really slow too(arount episode 100) and it is noisy(unstable) plot(biased).

Gamma = 0.95; it is starting to converge really soon, and compared to the last plot
,
it converges really fast (around episode 50).

by increasing the gamma, the variance is increasing as well and we will have unbiased expected return

Qn 2. ACTOR CRITIC [35 Marks]

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.

$\nabla_\theta J(\theta) = \sum_j$ **\ where $G^j_{t:t+1} = R_t$ is the truncated one-step return computed starting from the current**

$\sum_t (G^j_{t:t+1}$ $+ \gamma V(s^j_{t+1})$

$- V(s^j_t)$

$)\nabla_\theta ln \pi_\theta(a^j_t|s^j_t$

$)$

**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.

In [12]:

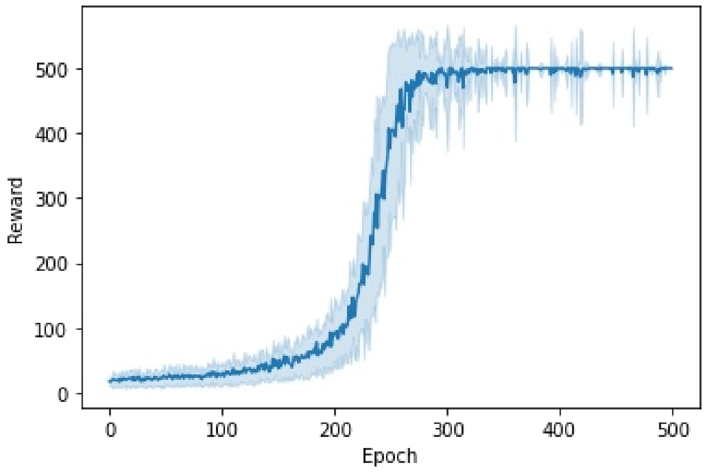
```
# Insert your code and run this cell
class ActorCriticAgent(BaseAgent):
    """ A2C Agent: Actor-Critic
        Here we try to FURTHER reduce the variance via bootstrapping.
    """

    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!
        """
        # ======================================================================

          # YOUR CODE HERE !
        total_rewards = np.zeros((n_episodes))
        policy_loss = 0.0
        value_loss = 0.0
        for i in range(n_episodes):

            current_state = self.env.reset()
            done = False
            episodes = []
            while not done:

                action_probability = self.policy_model.forward(torch.FloatTensor(current
_state).unsqueeze(0))

                action = np.random.choice(np.array([0,1]), p = action_probability.data.n
umpy()[0])
                prev_state = current_state
                current_state, reward, done, extra = self.env.step(action)
                episodes.append((prev_state, action, reward))

            reward_batch = np.array([r for (s,a,r) in episodes])



            expected_return = self._make_returns(reward_batch)
            total_rewards[i] = sum(reward_batch)

            expected_returns_batch = torch.FloatTensor(expected_return)
            state_batch = torch.Tensor([s for (s,a,r) in episodes])
            action_batch = torch.Tensor([a for (s,a,r) in episodes])

            value = self.value_model.forward(state_batch)

            G = torch.zeros_like(value)
            G[-1] = reward_batch[-1]
            for t in reversed(range(len(reward_batch) -1)):
                G[t] = (reward_batch[t] + self.gamma * value[t+1]).detach()




            prediction_batch = self.policy_model.forward(state_batch)
            action_selected_batch = prediction_batch.gather(dim = 1, index = action_batc
h.long().view(-1,1)).squeeze()

            policy_loss += - torch.sum(torch.log(action_selected_batch) * (G - value.det
ach()))
            value_loss +=   torch.sum((G.detach() - value).pow(2))

        value_loss /= n_episodes





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

        self.policy_optimizer.zero_grad()
        policy_loss.backward()
        self.policy_optimizer.step()

        self.value_optimizer.zero_grad()
        value_loss.backward()
        self.value_optimizer.step()
        return total_rewards
```

In [13]:

```
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)
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:48: UserWarning: Creating a
tensor from a list of numpy.ndarrays is extremely slow. Please consider converting the li
st to a single numpy.ndarray with numpy.array() before converting to a tensor. (Triggered
internally at  ../torch/csrc/utils/tensor_new.cpp:201.)
```

```
Iteration 1/500: rewards 17.91 +/- 6.97
Iteration 2/500: rewards 17.25 +/- 5.96
Iteration 3/500: rewards 18.84 +/- 8.69
Iteration 4/500: rewards 20.47 +/- 13.35
Iteration 5/500: rewards 20.97 +/- 10.87
Iteration 6/500: rewards 19.44 +/- 8.52
Iteration 7/500: rewards 19.03 +/- 8.87
Iteration 8/500: rewards 18.91 +/- 9.76
Iteration 9/500: rewards 20.91 +/- 11.14
Iteration 10/500: rewards 17.16 +/- 6.27
Iteration 11/500: rewards 21.34 +/- 12.76
Iteration 12/500: rewards 22.22 +/- 12.34
Iteration 13/500: rewards 19.28 +/- 6.35
Iteration 14/500: rewards 22.16 +/- 9.25
Iteration 15/500: rewards 21.19 +/- 14.63
Iteration 16/500: rewards 20.0 +/- 8.7
Iteration 17/500: rewards 21.78 +/- 9.8
Iteration 18/500: rewards 22.69 +/- 13.07
Iteration 19/500: rewards 21.59 +/- 13.93
Iteration 20/500: rewards 22.19 +/- 11.55
Iteration 21/500: rewards 23.84 +/- 13.68
Iteration 22/500: rewards 18.91 +/- 6.68
Iteration 23/500: rewards 23.62 +/- 15.24
Iteration 24/500: rewards 18.97 +/- 10.01
Iteration 25/500: rewards 22.84 +/- 13.61
Iteration 26/500: rewards 21.56 +/- 9.71
Iteration 27/500: rewards 25.25 +/- 14.0
Iteration 28/500: rewards 20.12 +/- 9.49
Iteration 29/500: rewards 17.56 +/- 6.81
Iteration 30/500: rewards 21.44 +/- 11.61
Iteration 31/500: rewards 18.59 +/- 7.76
Iteration 32/500: rewards 19.22 +/- 9.34
Iteration 33/500: rewards 23.09 +/- 15.72
Iteration 34/500: rewards 20.53 +/- 9.3
Iteration 35/500: rewards 23.41 +/- 12.47
Iteration 36/500: rewards 23.41 +/- 13.81
Iteration 37/500: rewards 22.22 +/- 9.55
Iteration 38/500: rewards 20.19 +/- 9.55
Iteration 39/500: rewards 19.66 +/- 11.89
Iteration 40/500: rewards 21.19 +/- 10.57
Iteration 41/500: rewards 24.38 +/- 13.7
Iteration 42/500: rewards 24.25 +/- 13.83
Iteration 43/500: rewards 23.06 +/- 10.67
Iteration 44/500: rewards 19.88 +/- 8.21
Iteration 45/500: rewards 21.72 +/- 11.65
Iteration 46/500: rewards 22.81 +/- 10.61
Iteration 47/500: rewards 22.06 +/- 15.52
Iteration 48/500: rewards 23.0 +/- 11.15
Iteration 49/500: rewards 26.88 +/- 17.16
Iteration 50/500: rewards 24.06 +/- 15.3
Iteration 51/500: rewards 23.84 +/- 10.03
Iteration 52/500: rewards 22.66 +/- 8.23
Iteration 53/500: rewards 24.84 +/- 13.5
Iteration 54/500: rewards 28.59 +/- 16.94
Iteration 55/500: rewards 24.88 +/- 13.02
Iteration 56/500: rewards 26.19 +/- 14.95
Iteration 57/500: rewards 21.16 +/- 11.69
Iteration 58/500: rewards 25.06 +/- 13.91
Iteration 59/500: rewards 27.69 +/- 16.74
Iteration 60/500: rewards 24.16 +/- 11.65
Iteration 61/500: rewards 23.31 +/- 9.2
Iteration 62/500: rewards 27.84 +/- 18.14
Iteration 63/500: rewards 22.62 +/- 11.31
Iteration 64/500: rewards 24.0 +/- 12.32
Iteration 65/500: rewards 25.25 +/- 12.09
Iteration 66/500: rewards 28.53 +/- 18.05
Iteration 67/500: rewards 22.0 +/- 11.1
Iteration 68/500: rewards 27.16 +/- 15.97
Iteration 69/500: rewards 26.66 +/- 12.03
Iteration 70/500: rewards 26.19 +/- 14.19
Iteration 71/500: rewards 28.16 +/- 14.98
Iteration 72/500: rewards 22.94 +/- 9.72
Iteration 73/500: rewards 27.12 +/- 16.23
Iteration 74/500: rewards 26.06 +/- 13.53
Iteration 75/500: rewards 22.97 +/- 9.56
Iteration 76/500: rewards 25.56 +/- 12.61
Iteration 77/500: rewards 27.41 +/- 16.48
Iteration 78/500: rewards 25.66 +/- 13.46
Iteration 79/500: rewards 25.53 +/- 9.09
Iteration 80/500: rewards 26.47 +/- 14.64
Iteration 81/500: rewards 23.94 +/- 8.17
Iteration 82/500: rewards 25.94 +/- 15.96
Iteration 83/500: rewards 21.19 +/- 10.23
Iteration 84/500: rewards 25.31 +/- 13.38
Iteration 85/500: rewards 25.84 +/- 13.66
Iteration 86/500: rewards 25.06 +/- 10.34
Iteration 87/500: rewards 27.19 +/- 14.12
Iteration 88/500: rewards 29.97 +/- 17.86
Iteration 89/500: rewards 26.06 +/- 15.91
Iteration 90/500: rewards 30.78 +/- 15.46
Iteration 91/500: rewards 25.59 +/- 13.52
Iteration 92/500: rewards 31.94 +/- 13.74
Iteration 93/500: rewards 32.28 +/- 22.43
Iteration 94/500: rewards 27.94 +/- 15.46
Iteration 95/500: rewards 31.03 +/- 15.66
Iteration 96/500: rewards 24.19 +/- 11.62
Iteration 97/500: rewards 30.53 +/- 17.28
Iteration 98/500: rewards 27.03 +/- 14.4
Iteration 99/500: rewards 27.12 +/- 16.75
Iteration 100/500: rewards 26.44 +/- 15.98
Iteration 101/500: rewards 30.88 +/- 17.37
Iteration 102/500: rewards 29.16 +/- 15.27
Iteration 103/500: rewards 24.84 +/- 9.94
Iteration 104/500: rewards 30.56 +/- 17.46
Iteration 105/500: rewards 32.5 +/- 18.9
Iteration 106/500: rewards 30.25 +/- 15.81
Iteration 107/500: rewards 31.97 +/- 21.0
Iteration 108/500: rewards 27.97 +/- 10.45
Iteration 109/500: rewards 29.31 +/- 17.86
Iteration 110/500: rewards 31.34 +/- 18.83
Iteration 111/500: rewards 37.44 +/- 25.35
Iteration 112/500: rewards 28.59 +/- 14.69
Iteration 113/500: rewards 33.38 +/- 18.8
Iteration 114/500: rewards 29.88 +/- 18.81
Iteration 115/500: rewards 32.88 +/- 18.87
Iteration 116/500: rewards 29.03 +/- 13.47
Iteration 117/500: rewards 31.97 +/- 14.75
Iteration 118/500: rewards 33.09 +/- 19.06
Iteration 119/500: rewards 31.66 +/- 14.29
Iteration 120/500: rewards 35.97 +/- 18.22
Iteration 121/500: rewards 39.25 +/- 23.57
Iteration 122/500: rewards 31.25 +/- 16.76
Iteration 123/500: rewards 35.34 +/- 21.26
Iteration 124/500: rewards 36.91 +/- 21.34
Iteration 125/500: rewards 37.06 +/- 21.3
Iteration 126/500: rewards 37.91 +/- 18.02
Iteration 127/500: rewards 32.53 +/- 18.24
Iteration 128/500: rewards 31.5 +/- 18.93
Iteration 129/500: rewards 36.5 +/- 23.58
Iteration 130/500: rewards 35.88 +/- 18.44
Iteration 131/500: rewards 32.06 +/- 16.19
Iteration 132/500: rewards 34.31 +/- 17.72
Iteration 133/500: rewards 40.91 +/- 25.37
Iteration 134/500: rewards 34.84 +/- 12.31
Iteration 135/500: rewards 41.75 +/- 24.97
Iteration 136/500: rewards 44.59 +/- 22.08
Iteration 137/500: rewards 43.88 +/- 22.45
Iteration 138/500: rewards 36.09 +/- 18.25
Iteration 139/500: rewards 35.16 +/- 17.26
Iteration 140/500: rewards 46.25 +/- 24.28
Iteration 141/500: rewards 43.16 +/- 19.2
Iteration 142/500: rewards 37.72 +/- 20.81
Iteration 143/500: rewards 32.56 +/- 14.51
Iteration 144/500: rewards 40.22 +/- 20.49
Iteration 145/500: rewards 38.72 +/- 19.77
Iteration 146/500: rewards 53.06 +/- 28.66
Iteration 147/500: rewards 40.75 +/- 20.07
Iteration 148/500: rewards 51.12 +/- 22.06
Iteration 149/500: rewards 42.84 +/- 22.19
Iteration 150/500: rewards 47.59 +/- 29.21
Iteration 151/500: rewards 37.56 +/- 19.42
Iteration 152/500: rewards 50.59 +/- 28.12
Iteration 153/500: rewards 41.03 +/- 19.91
Iteration 154/500: rewards 45.31 +/- 27.97
Iteration 155/500: rewards 48.28 +/- 25.48
Iteration 156/500: rewards 56.88 +/- 31.64
Iteration 157/500: rewards 43.78 +/- 25.35
Iteration 158/500: rewards 49.25 +/- 27.43
Iteration 159/500: rewards 55.53 +/- 30.4
Iteration 160/500: rewards 48.44 +/- 30.03
Iteration 161/500: rewards 45.19 +/- 20.2
Iteration 162/500: rewards 43.88 +/- 24.35
Iteration 163/500: rewards 39.5 +/- 16.89
Iteration 164/500: rewards 46.03 +/- 25.31
Iteration 165/500: rewards 55.44 +/- 24.04
Iteration 166/500: rewards 51.03 +/- 26.41
Iteration 167/500: rewards 56.72 +/- 30.95
Iteration 168/500: rewards 55.09 +/- 29.29
Iteration 169/500: rewards 47.22 +/- 25.23
Iteration 170/500: rewards 52.09 +/- 23.24
Iteration 171/500: rewards 55.0 +/- 26.27
Iteration 172/500: rewards 58.22 +/- 32.15
Iteration 173/500: rewards 53.94 +/- 32.83
Iteration 174/500: rewards 53.47 +/- 30.14
Iteration 175/500: rewards 54.66 +/- 29.16
Iteration 176/500: rewards 58.0 +/- 34.84
Iteration 177/500: rewards 63.06 +/- 41.3
Iteration 178/500: rewards 53.97 +/- 31.93
Iteration 179/500: rewards 63.75 +/- 36.7
Iteration 180/500: rewards 62.84 +/- 29.8
Iteration 181/500: rewards 60.31 +/- 30.45
Iteration 182/500: rewards 64.0 +/- 35.27
Iteration 183/500: rewards 68.09 +/- 33.74
Iteration 184/500: rewards 61.19 +/- 35.25
Iteration 185/500: rewards 77.5 +/- 37.03
Iteration 186/500: rewards 58.03 +/- 30.11
Iteration 187/500: rewards 63.97 +/- 34.14
Iteration 188/500: rewards 65.5 +/- 27.74
Iteration 189/500: rewards 60.75 +/- 31.28
Iteration 190/500: rewards 64.22 +/- 40.46
Iteration 191/500: rewards 75.5 +/- 33.57
Iteration 192/500: rewards 68.19 +/- 35.46
Iteration 193/500: rewards 75.62 +/- 45.24
Iteration 194/500: rewards 79.22 +/- 43.84
Iteration 195/500: rewards 78.0 +/- 36.32
Iteration 196/500: rewards 72.0 +/- 39.9
Iteration 197/500: rewards 93.19 +/- 52.33
Iteration 198/500: rewards 71.25 +/- 38.29
Iteration 199/500: rewards 82.41 +/- 44.36
Iteration 200/500: rewards 90.34 +/- 39.82
Iteration 201/500: rewards 93.03 +/- 53.27
Iteration 202/500: rewards 88.97 +/- 50.77
Iteration 203/500: rewards 94.88 +/- 39.31
Iteration 204/500: rewards 83.31 +/- 44.71
Iteration 205/500: rewards 87.06 +/- 45.96
Iteration 206/500: rewards 100.72 +/- 41.63
Iteration 207/500: rewards 95.91 +/- 56.77
Iteration 208/500: rewards 105.75 +/- 53.77
Iteration 209/500: rewards 90.53 +/- 46.17
Iteration 210/500: rewards 105.06 +/- 41.38
Iteration 211/500: rewards 102.81 +/- 53.41
Iteration 212/500: rewards 102.94 +/- 56.36
Iteration 213/500: rewards 111.94 +/- 53.13
Iteration 214/500: rewards 133.81 +/- 68.49
Iteration 215/500: rewards 114.03 +/- 68.44
Iteration 216/500: rewards 124.62 +/- 59.54
Iteration 217/500: rewards 147.06 +/- 59.59
Iteration 218/500: rewards 117.25 +/- 62.42
Iteration 219/500: rewards 122.72 +/- 70.61
Iteration 220/500: rewards 142.5 +/- 67.56
Iteration 221/500: rewards 147.91 +/- 96.87
Iteration 222/500: rewards 153.16 +/- 84.87
Iteration 223/500: rewards 167.59 +/- 70.56
Iteration 224/500: rewards 168.03 +/- 66.91
Iteration 225/500: rewards 174.97 +/- 81.71
Iteration 226/500: rewards 198.66 +/- 96.42
Iteration 227/500: rewards 167.12 +/- 112.73
Iteration 228/500: rewards 194.41 +/- 79.61
Iteration 229/500: rewards 185.62 +/- 107.68
Iteration 230/500: rewards 192.84 +/- 83.61
Iteration 231/500: rewards 182.28 +/- 113.48
Iteration 232/500: rewards 207.78 +/- 114.13
Iteration 233/500: rewards 241.75 +/- 121.0
Iteration 234/500: rewards 237.62 +/- 114.28
Iteration 235/500: rewards 229.03 +/- 107.71
Iteration 236/500: rewards 277.62 +/- 141.97
Iteration 237/500: rewards 256.78 +/- 144.61
Iteration 238/500: rewards 306.0 +/- 148.97
Iteration 239/500: rewards 253.91 +/- 102.62
Iteration 240/500: rewards 303.91 +/- 148.52
Iteration 241/500: rewards 300.84 +/- 129.49
Iteration 242/500: rewards 302.31 +/- 138.93
Iteration 243/500: rewards 319.12 +/- 136.46
Iteration 244/500: rewards 342.0 +/- 113.21
Iteration 245/500: rewards 298.84 +/- 145.7
Iteration 246/500: rewards 335.88 +/- 127.77
Iteration 247/500: rewards 348.62 +/- 137.88
Iteration 248/500: rewards 364.94 +/- 150.48
Iteration 249/500: rewards 406.78 +/- 124.65
Iteration 250/500: rewards 376.03 +/- 154.96
Iteration 251/500: rewards 403.34 +/- 132.86
Iteration 252/500: rewards 405.44 +/- 133.12
Iteration 253/500: rewards 403.72 +/- 130.35
Iteration 254/500: rewards 404.56 +/- 129.43
Iteration 255/500: rewards 394.78 +/- 141.16
Iteration 256/500: rewards 426.34 +/- 122.06
Iteration 257/500: rewards 443.19 +/- 114.36
Iteration 258/500: rewards 417.69 +/- 123.39
Iteration 259/500: rewards 421.88 +/- 109.76
Iteration 260/500: rewards 467.91 +/- 78.54
Iteration 261/500: rewards 408.62 +/- 137.85
Iteration 262/500: rewards 437.91 +/- 121.59
Iteration 263/500: rewards 436.75 +/- 109.52
Iteration 264/500: rewards 478.56 +/- 49.39
Iteration 265/500: rewards 465.72 +/- 76.2
Iteration 266/500: rewards 468.16 +/- 80.23
Iteration 267/500: rewards 473.81 +/- 62.4
Iteration 268/500: rewards 433.0 +/- 109.35
Iteration 269/500: rewards 480.91 +/- 74.69
Iteration 270/500: rewards 455.59 +/- 84.74
Iteration 271/500: rewards 478.97 +/- 57.95
Iteration 272/500: rewards 469.12 +/- 76.66
Iteration 273/500: rewards 472.28 +/- 81.02
Iteration 274/500: rewards 490.19 +/- 54.63
Iteration 275/500: rewards 472.31 +/- 86.44
Iteration 276/500: rewards 494.91 +/- 21.95
Iteration 277/500: rewards 475.12 +/- 81.8
Iteration 278/500: rewards 489.0 +/- 61.07
Iteration 279/500: rewards 481.66 +/- 70.24
Iteration 280/500: rewards 492.66 +/- 38.79
Iteration 281/500: rewards 489.94 +/- 25.8
Iteration 282/500: rewards 491.09 +/- 27.76
Iteration 283/500: rewards 484.56 +/- 35.4
Iteration 284/500: rewards 482.47 +/- 44.68
Iteration 285/500: rewards 478.94 +/- 70.01
Iteration 286/500: rewards 478.25 +/- 69.76
Iteration 287/500: rewards 493.16 +/- 38.1
Iteration 288/500: rewards 475.25 +/- 67.55
Iteration 289/500: rewards 489.62 +/- 57.77
Iteration 290/500: rewards 487.84 +/- 47.13
Iteration 291/500: rewards 476.16 +/- 80.98
Iteration 292/500: rewards 494.72 +/- 17.02
Iteration 293/500: rewards 492.25 +/- 32.28
Iteration 294/500: rewards 498.69 +/- 7.31
Iteration 295/500: rewards 492.88 +/- 27.88
Iteration 296/500: rewards 500.0 +/- 0.0
Iteration 297/500: rewards 494.59 +/- 30.1
Iteration 298/500: rewards 490.31 +/- 53.94
Iteration 299/500: rewards 492.12 +/- 37.61
Iteration 300/500: rewards 484.06 +/- 61.73
Iteration 301/500: rewards 470.25 +/- 95.43
Iteration 302/500: rewards 489.88 +/- 54.61
Iteration 303/500: rewards 500.0 +/- 0.0
Iteration 304/500: rewards 489.78 +/- 41.18
Iteration 305/500: rewards 500.0 +/- 0.0
Iteration 306/500: rewards 498.0 +/- 11.14
Iteration 307/500: rewards 491.88 +/- 37.97
Iteration 308/500: rewards 493.47 +/- 31.41
Iteration 309/500: rewards 492.19 +/- 43.5
Iteration 310/500: rewards 494.47 +/- 23.19
Iteration 311/500: rewards 478.69 +/- 70.26
Iteration 312/500: rewards 490.38 +/- 53.59
Iteration 313/500: rewards 488.84 +/- 62.12
Iteration 314/500: rewards 500.0 +/- 0.0
Iteration 315/500: rewards 469.72 +/- 91.96
Iteration 316/500: rewards 494.5 +/- 23.97
Iteration 317/500: rewards 496.09 +/- 21.75
Iteration 318/500: rewards 489.25 +/- 43.1
Iteration 319/500: rewards 497.75 +/- 12.53
Iteration 320/500: rewards 496.41 +/- 20.01
Iteration 321/500: rewards 492.72 +/- 40.54
Iteration 322/500: rewards 500.0 +/- 0.0
Iteration 323/500: rewards 498.59 +/- 5.48
Iteration 324/500: rewards 500.0 +/- 0.0
Iteration 325/500: rewards 494.22 +/- 32.19
Iteration 326/500: rewards 497.22 +/- 15.49
Iteration 327/500: rewards 498.94 +/- 5.92
Iteration 328/500: rewards 489.69 +/- 55.64
Iteration 329/500: rewards 493.75 +/- 34.8
Iteration 330/500: rewards 498.22 +/- 9.92
Iteration 331/500: rewards 493.31 +/- 37.23
Iteration 332/500: rewards 498.16 +/- 10.27
Iteration 333/500: rewards 496.66 +/- 18.62
Iteration 334/500: rewards 493.97 +/- 33.58
Iteration 335/500: rewards 491.91 +/- 45.06
Iteration 336/500: rewards 500.0 +/- 0.0
Iteration 337/500: rewards 500.0 +/- 0.0
Iteration 338/500: rewards 500.0 +/- 0.0
Iteration 339/500: rewards 500.0 +/- 0.0
Iteration 340/500: rewards 492.56 +/- 41.41
Iteration 341/500: rewards 500.0 +/- 0.0
Iteration 342/500: rewards 500.0 +/- 0.0
Iteration 343/500: rewards 496.88 +/- 17.4
Iteration 344/500: rewards 500.0 +/- 0.0
Iteration 345/500: rewards 495.97 +/- 22.45
Iteration 346/500: rewards 500.0 +/- 0.0
Iteration 347/500: rewards 500.0 +/- 0.0
Iteration 348/500: rewards 497.47 +/- 14.09
Iteration 349/500: rewards 500.0 +/- 0.0
Iteration 350/500: rewards 496.09 +/- 21.75
Iteration 351/500: rewards 500.0 +/- 0.0
Iteration 352/500: rewards 494.19 +/- 32.36
Iteration 353/500: rewards 493.56 +/- 35.84
Iteration 354/500: rewards 500.0 +/- 0.0
Iteration 355/500: rewards 500.0 +/- 0.0
Iteration 356/500: rewards 500.0 +/- 0.0
Iteration 357/500: rewards 500.0 +/- 0.0
Iteration 358/500: rewards 500.0 +/- 0.0
Iteration 359/500: rewards 493.62 +/- 35.49
Iteration 360/500: rewards 500.0 +/- 0.0
Iteration 361/500: rewards 477.56 +/- 87.69
Iteration 362/500: rewards 495.47 +/- 24.7
Iteration 363/500: rewards 500.0 +/- 0.0
Iteration 364/500: rewards 500.0 +/- 0.0
Iteration 365/500: rewards 500.0 +/- 0.0
Iteration 366/500: rewards 494.44 +/- 30.97
Iteration 367/500: rewards 500.0 +/- 0.0
Iteration 368/500: rewards 500.0 +/- 0.0
Iteration 369/500: rewards 498.84 +/- 6.44
Iteration 370/500: rewards 500.0 +/- 0.0
Iteration 371/500: rewards 491.47 +/- 47.5
Iteration 372/500: rewards 491.06 +/- 35.09
Iteration 373/500: rewards 500.0 +/- 0.0
Iteration 374/500: rewards 500.0 +/- 0.0
Iteration 375/500: rewards 500.0 +/- 0.0
Iteration 376/500: rewards 500.0 +/- 0.0
Iteration 377/500: rewards 500.0 +/- 0.0
Iteration 378/500: rewards 500.0 +/- 0.0
Iteration 379/500: rewards 500.0 +/- 0.0
Iteration 380/500: rewards 500.0 +/- 0.0
Iteration 381/500: rewards 500.0 +/- 0.0
Iteration 382/500: rewards 500.0 +/- 0.0
Iteration 383/500: rewards 498.22 +/- 9.92
Iteration 384/500: rewards 498.0 +/- 11.14
Iteration 385/500: rewards 498.25 +/- 9.74
Iteration 386/500: rewards 500.0 +/- 0.0
Iteration 387/500: rewards 500.0 +/- 0.0
Iteration 388/500: rewards 500.0 +/- 0.0
Iteration 389/500: rewards 500.0 +/- 0.0
Iteration 390/500: rewards 500.0 +/- 0.0
Iteration 391/500: rewards 500.0 +/- 0.0
Iteration 392/500: rewards 500.0 +/- 0.0
Iteration 393/500: rewards 489.69 +/- 57.42
Iteration 394/500: rewards 500.0 +/- 0.0
Iteration 395/500: rewards 500.0 +/- 0.0
Iteration 396/500: rewards 492.38 +/- 42.45
Iteration 397/500: rewards 500.0 +/- 0.0
Iteration 398/500: rewards 495.84 +/- 23.14
Iteration 399/500: rewards 500.0 +/- 0.0
Iteration 400/500: rewards 500.0 +/- 0.0
Iteration 401/500: rewards 500.0 +/- 0.0
Iteration 402/500: rewards 498.84 +/- 6.44
Iteration 403/500: rewards 500.0 +/- 0.0
Iteration 404/500: rewards 500.0 +/- 0.0
Iteration 405/500: rewards 498.56 +/- 8.0
Iteration 406/500: rewards 497.75 +/- 12.53
Iteration 407/500: rewards 495.41 +/- 25.58
Iteration 408/500: rewards 500.0 +/- 0.0
Iteration 409/500: rewards 500.0 +/- 0.0
Iteration 410/500: rewards 500.0 +/- 0.0
Iteration 411/500: rewards 491.03 +/- 49.94
Iteration 412/500: rewards 500.0 +/- 0.0
Iteration 413/500: rewards 500.0 +/- 0.0
Iteration 414/500: rewards 500.0 +/- 0.0
Iteration 415/500: rewards 488.22 +/- 58.35
Iteration 416/500: rewards 500.0 +/- 0.0
Iteration 417/500: rewards 497.44 +/- 14.27
Iteration 418/500: rewards 498.44 +/- 8.7
Iteration 419/500: rewards 486.94 +/- 72.73
Iteration 420/500: rewards 488.09 +/- 66.29
Iteration 421/500: rewards 496.88 +/- 17.4
Iteration 422/500: rewards 500.0 +/- 0.0
Iteration 423/500: rewards 499.25 +/- 4.18
Iteration 424/500: rewards 497.47 +/- 14.09
Iteration 425/500: rewards 498.5 +/- 8.35
Iteration 426/500: rewards 498.88 +/- 6.26
Iteration 427/500: rewards 500.0 +/- 0.0
Iteration 428/500: rewards 500.0 +/- 0.0
Iteration 429/500: rewards 500.0 +/- 0.0
Iteration 430/500: rewards 500.0 +/- 0.0
Iteration 431/500: rewards 500.0 +/- 0.0
Iteration 432/500: rewards 500.0 +/- 0.0
Iteration 433/500: rewards 499.09 +/- 5.05
Iteration 434/500: rewards 500.0 +/- 0.0
Iteration 435/500: rewards 498.38 +/- 9.05
Iteration 436/500: rewards 499.34 +/- 3.65
Iteration 437/500: rewards 498.0 +/- 9.65
Iteration 438/500: rewards 500.0 +/- 0.0
Iteration 439/500: rewards 500.0 +/- 0.0
Iteration 440/500: rewards 500.0 +/- 0.0
Iteration 441/500: rewards 500.0 +/- 0.0
Iteration 442/500: rewards 500.0 +/- 0.0
Iteration 443/500: rewards 500.0 +/- 0.0
Iteration 444/500: rewards 500.0 +/- 0.0
Iteration 445/500: rewards 500.0 +/- 0.0
Iteration 446/500: rewards 499.62 +/- 2.09
Iteration 447/500: rewards 500.0 +/- 0.0
Iteration 448/500: rewards 491.12 +/- 49.41
Iteration 449/500: rewards 500.0 +/- 0.0
Iteration 450/500: rewards 500.0 +/- 0.0
Iteration 451/500: rewards 500.0 +/- 0.0
Iteration 452/500: rewards 500.0 +/- 0.0
Iteration 453/500: rewards 500.0 +/- 0.0
Iteration 454/500: rewards 500.0 +/- 0.0
Iteration 455/500: rewards 500.0 +/- 0.0
Iteration 456/500: rewards 493.38 +/- 36.89
Iteration 457/500: rewards 500.0 +/- 0.0
Iteration 458/500: rewards 500.0 +/- 0.0
Iteration 459/500: rewards 500.0 +/- 0.0
Iteration 460/500: rewards 499.94 +/- 0.35
Iteration 461/500: rewards 500.0 +/- 0.0
Iteration 462/500: rewards 498.62 +/- 7.66
Iteration 463/500: rewards 500.0 +/- 0.0
Iteration 464/500: rewards 500.0 +/- 0.0
Iteration 465/500: rewards 500.0 +/- 0.0
Iteration 466/500: rewards 486.59 +/- 74.64
Iteration 467/500: rewards 500.0 +/- 0.0
Iteration 468/500: rewards 496.81 +/- 17.75
Iteration 469/500: rewards 500.0 +/- 0.0
Iteration 470/500: rewards 500.0 +/- 0.0
Iteration 471/500: rewards 493.91 +/- 33.93
Iteration 472/500: rewards 499.75 +/- 1.39
Iteration 473/500: rewards 500.0 +/- 0.0
Iteration 474/500: rewards 500.0 +/- 0.0
Iteration 475/500: rewards 500.0 +/- 0.0
Iteration 476/500: rewards 500.0 +/- 0.0
Iteration 477/500: rewards 500.0 +/- 0.0
Iteration 478/500: rewards 489.62 +/- 57.77
Iteration 479/500: rewards 500.0 +/- 0.0
Iteration 480/500: rewards 500.0 +/- 0.0
Iteration 481/500: rewards 500.0 +/- 0.0
Iteration 482/500: rewards 497.88 +/- 8.23
Iteration 483/500: rewards 500.0 +/- 0.0
Iteration 484/500: rewards 499.38 +/- 3.48
Iteration 485/500: rewards 500.0 +/- 0.0
Iteration 486/500: rewards 496.47 +/- 13.07
Iteration 487/500: rewards 495.59 +/- 18.22
Iteration 488/500: rewards 491.56 +/- 34.56
Iteration 489/500: rewards 499.03 +/- 5.39
Iteration 490/500: rewards 497.06 +/- 16.36
Iteration 491/500: rewards 500.0 +/- 0.0
Iteration 492/500: rewards 500.0 +/- 0.0
Iteration 493/500: rewards 500.0 +/- 0.0
Iteration 494/500: rewards 498.62 +/- 7.66
Iteration 495/500: rewards 500.0 +/- 0.0
Iteration 496/500: rewards 500.0 +/- 0.0
Iteration 497/500: rewards 500.0 +/- 0.0
Iteration 498/500: rewards 500.0 +/- 0.0
Iteration 499/500: rewards 500.0 +/- 0.0
Iteration 500/500: rewards 500.0 +/- 0.0
```



In [14]:

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

```
Reward: 500.0
```

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]

```
In the REINFORCE algorithm with state value function as a baseline,
we use discounted return(expected return) generated from
the current trajectory as our target(like monte Carlo).
So it is an offline unbiased algorithm, but in the ACTOR-CRITIC algorithm,
 we use the bootstrapping estimate as our target(like TD learning),
which updates every step using predictions of future return A critic is just an ob
server that provides feedback to an actor.
And also, it has an impact on a biased-variance trade-off.
Actor-critic helps decrease the variance, which also improves the performance.
To conclude, we can see REINFORCE with baseline as a version
 of actor-critic algorithm where the critic is the monte Carlo estimator
```

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]

```
In the actor-critic algorithm, we have an actor which
defines the policy and a critic which provides a more
reduced variance reward signal to update the actor.
Variance can be subtracted from a Monte-Carlo sample using
a more stable learned value function V(s) in the critic.
This value function is typically a neural network and can
be learned using either Monte-Carlo sampling or Temporal
difference (TD) learning.
So this value function can be biased if we use temporal
difference. So as we know, variance and biased are
difference. So as we know, variance and biased are
inversely related, so by having a lower variance, we have a
higher bias.

Also, how does it balance between variance and bias?
Increasing the reward step gets more similar to
monte-Carlo, so we would estimate the G_t with higher
accuracy, so it decreases the bias but increases the
variance on the other hand. So by deciding how many steps
we have in our critic, we can handle the variance-bias trade-off.

Also, Discount Factor Ensure that this captures how far
ahead agents should be predicting rewards for environments
where agents need to think thousands of steps into the
future, between pure TD learning and pure Monte-Carlo
sampling using a lambda parameter. By setting lambda to 0,
the algorithm reduces to TD learning, while setting it to 1
produces Monte-Carlo sampling.
```

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.

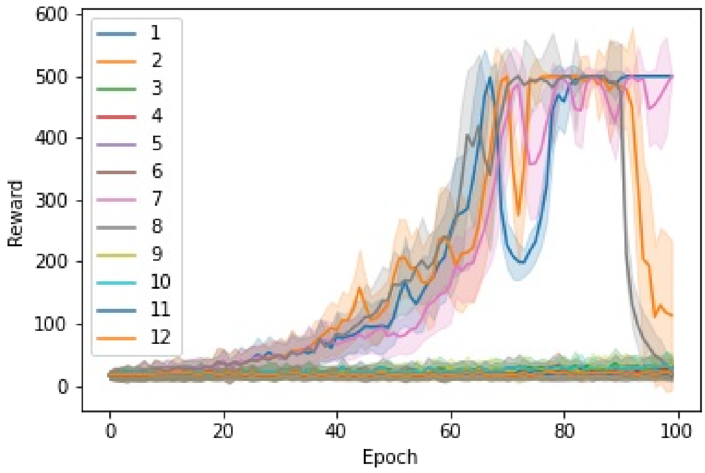
In [18]:

```
# Insert your code here to search for best hyper-parameters
Gamma = [1, 0.99]
Policy_network = [1e-2, 1e-3, 1e-4]
Value_network = [1e-2, 1e-3]
N_episodes = 50
N_iterations = 750
for i in range(len(Gamma)):
    for j in range(len(Policy_network)):
        for k in range(len(Value_network)):
            config = {
                'env_id': 'CartPole-v1',
                'seed': 8953,
                'gamma': Gamma[i],
                'policy_layers': [16, 8],
                'policy_learning_rate': Policy_network[j],
                'use_baseline': True,
                'value_layers': [16, 8, 8],
                'value_learning_rate': Value_network[k],
                    }
            agent = ActorCriticAgent(config)
            ActorCritic_rewards = agent.train(n_episodes=50, n_iterations=100)

plt.legend(labels=[1,2,3,4,5,6,7,8,9,10,11,12])

#(1, 1e-2, 1e-2) --- (1, 1e-2, 1e-3) --- (1, 1e-3, 1e-2) --- (1, 1e-3, 1e-3)
#(1, 1e-4, 1e-2) --- (1, 1e-4, 1e-3) --- (0.99, 1e-2, 1e-2) --- (0.99, 1e-2, 1e-3)
#(0.99, 1e-3, 1e-2) --- (0.99, 1e-3, 1e-3) --- (0.99, 1e-4, 1e-2) --- (0.99, 1e-4, 1e-3)
```

```
Iteration 1/100: rewards 17.8 +/- 6.91
Iteration 2/100: rewards 19.86 +/- 9.96
Iteration 3/100: rewards 17.78 +/- 7.26
Iteration 4/100: rewards 19.92 +/- 14.35
Iteration 5/100: rewards 20.82 +/- 9.36
Iteration 6/100: rewards 22.1 +/- 12.22
Iteration 7/100: rewards 25.1 +/- 15.89
Iteration 8/100: rewards 22.36 +/- 10.46
Iteration 9/100: rewards 26.64 +/- 15.97
Iteration 10/100: rewards 24.9 +/- 10.85
Iteration 11/100: rewards 27.4 +/- 15.11
Iteration 12/100: rewards 25.36 +/- 14.72
Iteration 13/100: rewards 23.0 +/- 12.22
Iteration 14/100: rewards 28.2 +/- 13.55
Iteration 15/100: rewards 27.3 +/- 15.19
Iteration 16/100: rewards 28.12 +/- 16.14
Iteration 17/100: rewards 30.34 +/- 15.64
Iteration 18/100: rewards 27.88 +/- 14.27
Iteration 19/100: rewards 29.76 +/- 14.35
Iteration 20/100: rewards 31.36 +/- 16.38
Iteration 21/100: rewards 39.6 +/- 19.74
Iteration 22/100: rewards 30.32 +/- 13.43
Iteration 23/100: rewards 33.56 +/- 15.85
Iteration 24/100: rewards 39.12 +/- 23.75
Iteration 25/100: rewards 37.96 +/- 19.38
Iteration 26/100: rewards 41.96 +/- 22.28
Iteration 27/100: rewards 50.1 +/- 27.73
Iteration 28/100: rewards 46.0 +/- 22.2
Iteration 29/100: rewards 53.84 +/- 26.66
Iteration 30/100: rewards 49.36 +/- 22.9
Iteration 31/100: rewards 44.6 +/- 25.75
Iteration 32/100: rewards 46.1 +/- 21.58
Iteration 33/100: rewards 54.06 +/- 26.69
Iteration 34/100: rewards 54.28 +/- 26.28
Iteration 35/100: rewards 56.06 +/- 21.02
Iteration 36/100: rewards 52.12 +/- 26.52
Iteration 37/100: rewards 56.92 +/- 26.57
Iteration 38/100: rewards 72.2 +/- 33.67
Iteration 39/100: rewards 71.18 +/- 32.29
Iteration 40/100: rewards 61.04 +/- 16.8
Iteration 41/100: rewards 78.96 +/- 38.27
Iteration 42/100: rewards 76.36 +/- 27.35
Iteration 43/100: rewards 79.88 +/- 26.58
Iteration 44/100: rewards 82.2 +/- 27.39
Iteration 45/100: rewards 87.4 +/- 33.3
Iteration 46/100: rewards 95.96 +/- 42.66
Iteration 47/100: rewards 94.84 +/- 35.89
Iteration 48/100: rewards 95.16 +/- 41.59
Iteration 49/100: rewards 96.92 +/- 44.83
Iteration 50/100: rewards 93.8 +/- 32.52
Iteration 51/100: rewards 109.28 +/- 41.18
Iteration 52/100: rewards 142.88 +/- 49.4
Iteration 53/100: rewards 169.26 +/- 71.14
Iteration 54/100: rewards 143.74 +/- 51.97
Iteration 55/100: rewards 131.86 +/- 45.66
Iteration 56/100: rewards 148.56 +/- 60.39
Iteration 57/100: rewards 165.12 +/- 43.74
Iteration 58/100: rewards 176.0 +/- 69.05
Iteration 59/100: rewards 197.1 +/- 64.64
Iteration 60/100: rewards 208.06 +/- 71.03
Iteration 61/100: rewards 248.14 +/- 73.09
Iteration 62/100: rewards 273.84 +/- 94.01
Iteration 63/100: rewards 279.2 +/- 91.0
Iteration 64/100: rewards 285.88 +/- 96.12
Iteration 65/100: rewards 332.44 +/- 129.83
Iteration 66/100: rewards 388.46 +/- 129.84
Iteration 67/100: rewards 466.22 +/- 76.11
Iteration 68/100: rewards 497.9 +/- 14.7
Iteration 69/100: rewards 446.58 +/- 74.76
Iteration 70/100: rewards 284.88 +/- 52.55
Iteration 71/100: rewards 226.7 +/- 36.69
Iteration 72/100: rewards 209.46 +/- 36.62
Iteration 73/100: rewards 200.3 +/- 28.65
Iteration 74/100: rewards 198.88 +/- 20.86
Iteration 75/100: rewards 217.68 +/- 30.28
Iteration 76/100: rewards 232.56 +/- 29.28
Iteration 77/100: rewards 271.88 +/- 45.54
Iteration 78/100: rewards 324.08 +/- 55.85
Iteration 79/100: rewards 446.34 +/- 63.58
Iteration 80/100: rewards 469.36 +/- 49.62
Iteration 81/100: rewards 458.34 +/- 50.03
Iteration 82/100: rewards 485.42 +/- 32.44
Iteration 83/100: rewards 494.56 +/- 21.44
Iteration 84/100: rewards 497.82 +/- 11.12
Iteration 85/100: rewards 500.0 +/- 0.0
Iteration 86/100: rewards 500.0 +/- 0.0
Iteration 87/100: rewards 498.0 +/- 10.42
Iteration 88/100: rewards 493.72 +/- 16.12
Iteration 89/100: rewards 490.76 +/- 16.49
Iteration 90/100: rewards 486.06 +/- 23.08
Iteration 91/100: rewards 495.98 +/- 11.98
Iteration 92/100: rewards 499.7 +/- 1.59
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
Iteration 1/100: rewards 17.8 +/- 6.91
Iteration 2/100: rewards 19.86 +/- 9.96
Iteration 3/100: rewards 17.78 +/- 7.26
Iteration 4/100: rewards 19.92 +/- 14.35
Iteration 5/100: rewards 20.82 +/- 9.36
Iteration 6/100: rewards 22.1 +/- 12.22
Iteration 7/100: rewards 25.1 +/- 15.89
Iteration 8/100: rewards 22.36 +/- 10.46
Iteration 9/100: rewards 27.14 +/- 14.57
Iteration 10/100: rewards 26.06 +/- 13.39
Iteration 11/100: rewards 23.62 +/- 9.82
Iteration 12/100: rewards 26.78 +/- 16.5
Iteration 13/100: rewards 30.2 +/- 17.55
Iteration 14/100: rewards 27.14 +/- 12.95
Iteration 15/100: rewards 26.12 +/- 15.29
Iteration 16/100: rewards 26.9 +/- 13.9
Iteration 17/100: rewards 27.34 +/- 14.59
Iteration 18/100: rewards 30.52 +/- 15.39
Iteration 19/100: rewards 28.94 +/- 14.54
Iteration 20/100: rewards 38.38 +/- 23.45
Iteration 21/100: rewards 35.88 +/- 22.93
Iteration 22/100: rewards 32.94 +/- 18.3
Iteration 23/100: rewards 38.24 +/- 22.43
Iteration 24/100: rewards 33.7 +/- 17.81
Iteration 25/100: rewards 37.8 +/- 21.25
Iteration 26/100: rewards 41.58 +/- 22.97
Iteration 27/100: rewards 40.06 +/- 23.02
Iteration 28/100: rewards 44.84 +/- 20.77
Iteration 29/100: rewards 45.76 +/- 27.51
Iteration 30/100: rewards 45.32 +/- 24.88
Iteration 31/100: rewards 47.04 +/- 27.96
Iteration 32/100: rewards 48.14 +/- 24.99
Iteration 33/100: rewards 56.38 +/- 34.14
Iteration 34/100: rewards 56.1 +/- 24.21
Iteration 35/100: rewards 58.76 +/- 23.23
Iteration 36/100: rewards 57.34 +/- 22.21
Iteration 37/100: rewards 69.74 +/- 35.09
Iteration 38/100: rewards 78.24 +/- 27.09
Iteration 39/100: rewards 84.62 +/- 39.15
Iteration 40/100: rewards 93.1 +/- 42.34
Iteration 41/100: rewards 90.34 +/- 39.51
Iteration 42/100: rewards 105.7 +/- 46.69
Iteration 43/100: rewards 106.14 +/- 40.81
Iteration 44/100: rewards 125.08 +/- 50.14
Iteration 45/100: rewards 158.28 +/- 61.35
Iteration 46/100: rewards 130.56 +/- 42.36
Iteration 47/100: rewards 105.98 +/- 37.14
Iteration 48/100: rewards 109.32 +/- 41.17
Iteration 49/100: rewards 119.42 +/- 40.54
Iteration 50/100: rewards 131.26 +/- 41.68
Iteration 51/100: rewards 174.0 +/- 59.74
Iteration 52/100: rewards 204.84 +/- 64.46
Iteration 53/100: rewards 206.62 +/- 61.48
Iteration 54/100: rewards 189.68 +/- 48.56
Iteration 55/100: rewards 190.42 +/- 65.25
Iteration 56/100: rewards 167.62 +/- 53.74
Iteration 57/100: rewards 166.08 +/- 74.56
Iteration 58/100: rewards 175.16 +/- 76.55
Iteration 59/100: rewards 235.52 +/- 82.58
Iteration 60/100: rewards 241.44 +/- 80.19
Iteration 61/100: rewards 223.78 +/- 84.18
Iteration 62/100: rewards 192.48 +/- 68.54
Iteration 63/100: rewards 213.26 +/- 92.74
Iteration 64/100: rewards 215.42 +/- 80.62
Iteration 65/100: rewards 222.54 +/- 86.37
Iteration 66/100: rewards 259.44 +/- 100.65
Iteration 67/100: rewards 321.6 +/- 123.8
Iteration 68/100: rewards 374.96 +/- 127.84
Iteration 69/100: rewards 445.1 +/- 100.68
Iteration 70/100: rewards 491.72 +/- 41.68
Iteration 71/100: rewards 499.3 +/- 4.9
Iteration 72/100: rewards 352.86 +/- 75.91
Iteration 73/100: rewards 275.44 +/- 80.01
Iteration 74/100: rewards 353.68 +/- 75.75
Iteration 75/100: rewards 496.08 +/- 13.54
Iteration 76/100: rewards 489.78 +/- 39.54
Iteration 77/100: rewards 498.4 +/- 11.2
Iteration 78/100: rewards 498.98 +/- 5.45
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 494.62 +/- 37.66
Iteration 85/100: rewards 491.76 +/- 42.02
Iteration 86/100: rewards 500.0 +/- 0.0
Iteration 87/100: rewards 486.76 +/- 57.61
Iteration 88/100: rewards 495.4 +/- 32.2
Iteration 89/100: rewards 477.16 +/- 82.34
Iteration 90/100: rewards 488.22 +/- 59.16
Iteration 91/100: rewards 483.18 +/- 68.99
Iteration 92/100: rewards 479.14 +/- 61.56
Iteration 93/100: rewards 450.26 +/- 127.39
Iteration 94/100: rewards 331.28 +/- 164.56
Iteration 95/100: rewards 204.12 +/- 154.88
Iteration 96/100: rewards 194.08 +/- 130.02
Iteration 97/100: rewards 111.64 +/- 112.34
Iteration 98/100: rewards 129.96 +/- 124.18
Iteration 99/100: rewards 117.84 +/- 126.41
Iteration 100/100: rewards 113.88 +/- 118.85
Iteration 1/100: rewards 17.8 +/- 6.91
Iteration 2/100: rewards 18.44 +/- 8.44
Iteration 3/100: rewards 19.42 +/- 11.75
Iteration 4/100: rewards 21.34 +/- 9.61
Iteration 5/100: rewards 18.28 +/- 9.04
Iteration 6/100: rewards 20.28 +/- 11.91
Iteration 7/100: rewards 18.04 +/- 8.19
Iteration 8/100: rewards 21.48 +/- 11.12
Iteration 9/100: rewards 20.92 +/- 8.89
Iteration 10/100: rewards 20.28 +/- 10.97
Iteration 11/100: rewards 20.56 +/- 11.02
Iteration 12/100: rewards 20.38 +/- 11.69
Iteration 13/100: rewards 21.14 +/- 9.44
Iteration 14/100: rewards 21.02 +/- 13.52
Iteration 15/100: rewards 22.18 +/- 10.63
Iteration 16/100: rewards 20.72 +/- 10.17
Iteration 17/100: rewards 23.18 +/- 14.78
Iteration 18/100: rewards 20.4 +/- 8.41
Iteration 19/100: rewards 19.68 +/- 10.34
Iteration 20/100: rewards 20.02 +/- 9.43
Iteration 21/100: rewards 19.4 +/- 6.23
Iteration 22/100: rewards 22.2 +/- 11.16
Iteration 23/100: rewards 19.22 +/- 6.54
Iteration 24/100: rewards 21.72 +/- 9.24
Iteration 25/100: rewards 20.58 +/- 10.06
Iteration 26/100: rewards 22.4 +/- 12.13
Iteration 27/100: rewards 20.06 +/- 9.94
Iteration 28/100: rewards 22.62 +/- 10.67
Iteration 29/100: rewards 23.48 +/- 13.5
Iteration 30/100: rewards 22.86 +/- 14.98
Iteration 31/100: rewards 20.62 +/- 9.15
Iteration 32/100: rewards 25.38 +/- 11.63
Iteration 33/100: rewards 21.1 +/- 10.41
Iteration 34/100: rewards 22.48 +/- 9.69
Iteration 35/100: rewards 22.34 +/- 10.36
Iteration 36/100: rewards 25.42 +/- 14.33
Iteration 37/100: rewards 21.82 +/- 10.39
Iteration 38/100: rewards 21.54 +/- 12.19
Iteration 39/100: rewards 21.02 +/- 8.84
Iteration 40/100: rewards 21.6 +/- 9.41
Iteration 41/100: rewards 20.3 +/- 8.3
Iteration 42/100: rewards 22.94 +/- 13.43
Iteration 43/100: rewards 19.98 +/- 8.4
Iteration 44/100: rewards 23.38 +/- 10.88
Iteration 45/100: rewards 21.76 +/- 9.13
Iteration 46/100: rewards 23.18 +/- 10.33
Iteration 47/100: rewards 22.26 +/- 11.46
Iteration 48/100: rewards 23.7 +/- 12.82
Iteration 49/100: rewards 22.98 +/- 13.47
Iteration 50/100: rewards 26.64 +/- 17.01
Iteration 51/100: rewards 22.92 +/- 10.86
Iteration 52/100: rewards 22.12 +/- 11.19
Iteration 53/100: rewards 22.82 +/- 15.97
Iteration 54/100: rewards 22.42 +/- 8.53
Iteration 55/100: rewards 25.4 +/- 14.21
Iteration 56/100: rewards 22.62 +/- 8.74
Iteration 57/100: rewards 20.84 +/- 7.84
Iteration 58/100: rewards 26.02 +/- 12.62
Iteration 59/100: rewards 27.96 +/- 19.85
Iteration 60/100: rewards 26.54 +/- 14.43
Iteration 61/100: rewards 21.86 +/- 12.04
Iteration 62/100: rewards 25.92 +/- 12.85
Iteration 63/100: rewards 26.9 +/- 14.71
Iteration 64/100: rewards 22.98 +/- 10.05
Iteration 65/100: rewards 25.76 +/- 12.13
Iteration 66/100: rewards 25.1 +/- 12.4
Iteration 67/100: rewards 26.4 +/- 14.27
Iteration 68/100: rewards 24.9 +/- 15.47
Iteration 69/100: rewards 27.72 +/- 17.47
Iteration 70/100: rewards 25.96 +/- 16.83
Iteration 71/100: rewards 25.52 +/- 11.34
Iteration 72/100: rewards 26.14 +/- 11.96
Iteration 73/100: rewards 24.84 +/- 11.47
Iteration 74/100: rewards 28.76 +/- 18.59
Iteration 75/100: rewards 30.26 +/- 17.11
Iteration 76/100: rewards 26.36 +/- 11.6
Iteration 77/100: rewards 29.34 +/- 15.15
Iteration 78/100: rewards 25.7 +/- 16.22
Iteration 79/100: rewards 24.6 +/- 14.29
Iteration 80/100: rewards 32.7 +/- 18.88
Iteration 81/100: rewards 28.66 +/- 16.13
Iteration 82/100: rewards 27.12 +/- 14.97
Iteration 83/100: rewards 32.9 +/- 22.51
Iteration 84/100: rewards 24.46 +/- 13.74
Iteration 85/100: rewards 32.24 +/- 14.71
Iteration 86/100: rewards 28.32 +/- 15.62
Iteration 87/100: rewards 29.6 +/- 19.26
Iteration 88/100: rewards 26.98 +/- 13.24
Iteration 89/100: rewards 29.34 +/- 19.59
Iteration 90/100: rewards 36.16 +/- 24.51
Iteration 91/100: rewards 30.34 +/- 18.82
Iteration 92/100: rewards 30.2 +/- 16.22
Iteration 93/100: rewards 31.08 +/- 17.99
Iteration 94/100: rewards 32.12 +/- 18.23
Iteration 95/100: rewards 26.24 +/- 15.12
Iteration 96/100: rewards 29.58 +/- 18.64
Iteration 97/100: rewards 33.52 +/- 17.82
Iteration 98/100: rewards 28.5 +/- 13.1
Iteration 99/100: rewards 34.54 +/- 21.61
Iteration 100/100: rewards 31.7 +/- 18.74
Iteration 1/100: rewards 17.8 +/- 6.91
Iteration 2/100: rewards 18.44 +/- 8.44
Iteration 3/100: rewards 19.42 +/- 11.75
Iteration 4/100: rewards 21.34 +/- 9.61
Iteration 5/100: rewards 18.28 +/- 9.04
Iteration 6/100: rewards 20.28 +/- 11.91
Iteration 7/100: rewards 18.04 +/- 8.19
Iteration 8/100: rewards 21.48 +/- 11.12
Iteration 9/100: rewards 20.92 +/- 8.89
Iteration 10/100: rewards 20.28 +/- 10.97
Iteration 11/100: rewards 20.56 +/- 11.02
Iteration 12/100: rewards 20.38 +/- 11.69
Iteration 13/100: rewards 21.14 +/- 9.44
Iteration 14/100: rewards 21.02 +/- 13.52
Iteration 15/100: rewards 22.18 +/- 10.63
Iteration 16/100: rewards 20.72 +/- 10.17
Iteration 17/100: rewards 23.18 +/- 14.78
Iteration 18/100: rewards 20.4 +/- 8.41
Iteration 19/100: rewards 19.68 +/- 10.34
Iteration 20/100: rewards 20.02 +/- 9.43
Iteration 21/100: rewards 19.4 +/- 6.23
Iteration 22/100: rewards 22.2 +/- 11.16
Iteration 23/100: rewards 19.22 +/- 6.54
Iteration 24/100: rewards 21.72 +/- 9.24
Iteration 25/100: rewards 20.58 +/- 10.06
Iteration 26/100: rewards 22.4 +/- 12.13
Iteration 27/100: rewards 20.06 +/- 9.94
Iteration 28/100: rewards 22.62 +/- 10.67
Iteration 29/100: rewards 23.48 +/- 13.5
Iteration 30/100: rewards 22.86 +/- 14.98
Iteration 31/100: rewards 20.62 +/- 9.15
Iteration 32/100: rewards 25.38 +/- 11.63
Iteration 33/100: rewards 21.1 +/- 10.41
Iteration 34/100: rewards 22.48 +/- 9.69
Iteration 35/100: rewards 22.34 +/- 10.36
Iteration 36/100: rewards 25.42 +/- 14.33
Iteration 37/100: rewards 21.82 +/- 10.39
Iteration 38/100: rewards 21.54 +/- 12.19
Iteration 39/100: rewards 21.54 +/- 10.4
Iteration 40/100: rewards 22.0 +/- 10.49
Iteration 41/100: rewards 22.84 +/- 10.52
Iteration 42/100: rewards 29.58 +/- 14.88
Iteration 43/100: rewards 22.02 +/- 11.4
Iteration 44/100: rewards 23.34 +/- 9.76
Iteration 45/100: rewards 25.08 +/- 14.79
Iteration 46/100: rewards 23.0 +/- 11.3
Iteration 47/100: rewards 23.1 +/- 12.79
Iteration 48/100: rewards 24.62 +/- 11.96
Iteration 49/100: rewards 29.7 +/- 15.64
Iteration 50/100: rewards 23.98 +/- 14.83
Iteration 51/100: rewards 21.9 +/- 11.24
Iteration 52/100: rewards 23.42 +/- 9.71
Iteration 53/100: rewards 22.68 +/- 10.55
Iteration 54/100: rewards 25.9 +/- 14.35
Iteration 55/100: rewards 23.16 +/- 13.25
Iteration 56/100: rewards 20.52 +/- 10.17
Iteration 57/100: rewards 27.76 +/- 18.26
Iteration 58/100: rewards 26.42 +/- 12.35
Iteration 59/100: rewards 24.14 +/- 15.42
Iteration 60/100: rewards 23.3 +/- 13.42
Iteration 61/100: rewards 23.16 +/- 12.89
Iteration 62/100: rewards 24.12 +/- 12.82
Iteration 63/100: rewards 25.3 +/- 14.79
Iteration 64/100: rewards 24.72 +/- 15.07
Iteration 65/100: rewards 25.9 +/- 14.5
Iteration 66/100: rewards 21.38 +/- 9.9
Iteration 67/100: rewards 26.96 +/- 13.17
Iteration 68/100: rewards 23.18 +/- 10.99
Iteration 69/100: rewards 25.46 +/- 12.55
Iteration 70/100: rewards 24.6 +/- 11.49
Iteration 71/100: rewards 28.12 +/- 19.27
Iteration 72/100: rewards 28.02 +/- 15.47
Iteration 73/100: rewards 30.36 +/- 18.75
Iteration 74/100: rewards 28.88 +/- 13.06
Iteration 75/100: rewards 24.24 +/- 13.26
Iteration 76/100: rewards 25.88 +/- 13.93
Iteration 77/100: rewards 23.84 +/- 13.57
Iteration 78/100: rewards 25.82 +/- 12.17
Iteration 79/100: rewards 29.58 +/- 18.93
Iteration 80/100: rewards 31.3 +/- 22.6
Iteration 81/100: rewards 26.4 +/- 15.26
Iteration 82/100: rewards 26.12 +/- 12.74
Iteration 83/100: rewards 28.14 +/- 15.73
Iteration 84/100: rewards 24.72 +/- 13.53
Iteration 85/100: rewards 24.6 +/- 10.25
Iteration 86/100: rewards 25.72 +/- 12.38
Iteration 87/100: rewards 26.26 +/- 14.94
Iteration 88/100: rewards 27.84 +/- 16.12
Iteration 89/100: rewards 26.24 +/- 13.85
Iteration 90/100: rewards 29.64 +/- 22.14
Iteration 91/100: rewards 27.94 +/- 12.0
Iteration 92/100: rewards 31.38 +/- 18.86
Iteration 93/100: rewards 29.2 +/- 17.17
Iteration 94/100: rewards 24.32 +/- 12.72
Iteration 95/100: rewards 31.1 +/- 15.25
Iteration 96/100: rewards 30.36 +/- 18.05
Iteration 97/100: rewards 29.56 +/- 16.5
Iteration 98/100: rewards 29.44 +/- 13.02
Iteration 99/100: rewards 28.14 +/- 15.34
Iteration 100/100: rewards 29.62 +/- 13.41
Iteration 1/100: rewards 17.8 +/- 6.91
Iteration 2/100: rewards 18.56 +/- 9.3
Iteration 3/100: rewards 18.18 +/- 9.35
Iteration 4/100: rewards 18.18 +/- 7.64
Iteration 5/100: rewards 17.4 +/- 8.27
Iteration 6/100: rewards 18.2 +/- 8.56
Iteration 7/100: rewards 17.06 +/- 8.04
Iteration 8/100: rewards 19.04 +/- 9.26
Iteration 9/100: rewards 19.22 +/- 11.09
Iteration 10/100: rewards 20.34 +/- 10.64
Iteration 11/100: rewards 19.46 +/- 9.32
Iteration 12/100: rewards 20.84 +/- 11.27
Iteration 13/100: rewards 20.52 +/- 11.57
Iteration 14/100: rewards 20.54 +/- 11.54
Iteration 15/100: rewards 20.82 +/- 11.28
Iteration 16/100: rewards 19.52 +/- 11.1
Iteration 17/100: rewards 18.8 +/- 9.53
Iteration 18/100: rewards 21.48 +/- 10.77
Iteration 19/100: rewards 21.5 +/- 10.5
Iteration 20/100: rewards 18.0 +/- 8.7
Iteration 21/100: rewards 19.28 +/- 9.1
Iteration 22/100: rewards 20.58 +/- 16.19
Iteration 23/100: rewards 19.14 +/- 7.75
Iteration 24/100: rewards 20.26 +/- 10.12
Iteration 25/100: rewards 17.7 +/- 6.38
Iteration 26/100: rewards 20.54 +/- 12.12
Iteration 27/100: rewards 18.24 +/- 8.09
Iteration 28/100: rewards 18.28 +/- 7.18
Iteration 29/100: rewards 18.28 +/- 6.99
Iteration 30/100: rewards 18.08 +/- 7.67
Iteration 31/100: rewards 19.64 +/- 8.65
Iteration 32/100: rewards 19.96 +/- 8.52
Iteration 33/100: rewards 20.22 +/- 10.6
Iteration 34/100: rewards 22.32 +/- 11.55
Iteration 35/100: rewards 17.38 +/- 8.66
Iteration 36/100: rewards 19.14 +/- 9.56
Iteration 37/100: rewards 20.48 +/- 8.96
Iteration 38/100: rewards 18.16 +/- 7.62
Iteration 39/100: rewards 18.62 +/- 6.66
Iteration 40/100: rewards 20.12 +/- 9.39
Iteration 41/100: rewards 19.96 +/- 9.97
Iteration 42/100: rewards 17.58 +/- 6.65
Iteration 43/100: rewards 20.04 +/- 10.36
Iteration 44/100: rewards 17.6 +/- 7.32
Iteration 45/100: rewards 19.64 +/- 10.05
Iteration 46/100: rewards 19.34 +/- 12.17
Iteration 47/100: rewards 17.76 +/- 7.42
Iteration 48/100: rewards 20.24 +/- 10.68
Iteration 49/100: rewards 21.38 +/- 9.21
Iteration 50/100: rewards 19.02 +/- 8.99
Iteration 51/100: rewards 18.74 +/- 7.92
Iteration 52/100: rewards 18.62 +/- 8.55
Iteration 53/100: rewards 19.56 +/- 12.44
Iteration 54/100: rewards 19.2 +/- 9.67
Iteration 55/100: rewards 17.36 +/- 7.52
Iteration 56/100: rewards 17.82 +/- 7.41
Iteration 57/100: rewards 19.9 +/- 9.6
Iteration 58/100: rewards 18.78 +/- 6.42
Iteration 59/100: rewards 20.04 +/- 10.56
Iteration 60/100: rewards 18.2 +/- 11.72
Iteration 61/100: rewards 18.78 +/- 7.84
Iteration 62/100: rewards 19.3 +/- 7.96
Iteration 63/100: rewards 19.5 +/- 9.41
Iteration 64/100: rewards 21.1 +/- 10.87
Iteration 65/100: rewards 18.38 +/- 7.41
Iteration 66/100: rewards 22.02 +/- 11.24
Iteration 67/100: rewards 19.12 +/- 8.33
Iteration 68/100: rewards 18.1 +/- 8.63
Iteration 69/100: rewards 19.04 +/- 8.95
Iteration 70/100: rewards 18.96 +/- 8.65
Iteration 71/100: rewards 18.64 +/- 9.11
Iteration 72/100: rewards 18.02 +/- 7.12
Iteration 73/100: rewards 17.94 +/- 7.66
Iteration 74/100: rewards 22.02 +/- 12.4
Iteration 75/100: rewards 16.9 +/- 6.28
Iteration 76/100: rewards 20.24 +/- 9.12
Iteration 77/100: rewards 19.3 +/- 10.26
Iteration 78/100: rewards 20.6 +/- 10.82
Iteration 79/100: rewards 19.58 +/- 10.82
Iteration 80/100: rewards 21.7 +/- 12.44
Iteration 81/100: rewards 20.18 +/- 8.5
Iteration 82/100: rewards 18.82 +/- 8.76
Iteration 83/100: rewards 21.46 +/- 11.25
Iteration 84/100: rewards 22.22 +/- 12.09
Iteration 85/100: rewards 20.66 +/- 9.8
Iteration 86/100: rewards 19.56 +/- 10.07
Iteration 87/100: rewards 20.36 +/- 8.82
Iteration 88/100: rewards 19.32 +/- 7.74
Iteration 89/100: rewards 18.94 +/- 9.07
Iteration 90/100: rewards 18.9 +/- 7.88
Iteration 91/100: rewards 20.26 +/- 11.7
Iteration 92/100: rewards 22.42 +/- 14.55
Iteration 93/100: rewards 18.34 +/- 8.29
Iteration 94/100: rewards 18.9 +/- 8.98
Iteration 95/100: rewards 19.64 +/- 7.8
Iteration 96/100: rewards 19.64 +/- 8.57
Iteration 97/100: rewards 18.86 +/- 9.09
Iteration 98/100: rewards 20.02 +/- 9.73
Iteration 99/100: rewards 19.06 +/- 10.8
Iteration 100/100: rewards 20.94 +/- 10.34
Iteration 1/100: rewards 17.8 +/- 6.91
Iteration 2/100: rewards 18.56 +/- 9.3
Iteration 3/100: rewards 18.18 +/- 9.35
Iteration 4/100: rewards 18.18 +/- 7.64
Iteration 5/100: rewards 17.4 +/- 8.27
Iteration 6/100: rewards 18.2 +/- 8.56
Iteration 7/100: rewards 17.06 +/- 8.04
Iteration 8/100: rewards 19.04 +/- 9.26
Iteration 9/100: rewards 19.22 +/- 11.09
Iteration 10/100: rewards 20.34 +/- 10.64
Iteration 11/100: rewards 19.46 +/- 9.32
Iteration 12/100: rewards 20.84 +/- 11.27
Iteration 13/100: rewards 20.52 +/- 11.57
Iteration 14/100: rewards 20.54 +/- 11.54
Iteration 15/100: rewards 20.82 +/- 11.28
Iteration 16/100: rewards 19.52 +/- 11.1
Iteration 17/100: rewards 18.8 +/- 9.53
Iteration 18/100: rewards 21.48 +/- 10.77
Iteration 19/100: rewards 21.5 +/- 10.5
Iteration 20/100: rewards 18.0 +/- 8.7
Iteration 21/100: rewards 19.28 +/- 9.1
Iteration 22/100: rewards 20.58 +/- 16.19
Iteration 23/100: rewards 19.14 +/- 7.75
Iteration 24/100: rewards 20.26 +/- 10.12
Iteration 25/100: rewards 17.7 +/- 6.38
Iteration 26/100: rewards 20.54 +/- 12.12
Iteration 27/100: rewards 18.24 +/- 8.09
Iteration 28/100: rewards 18.28 +/- 7.18
Iteration 29/100: rewards 18.28 +/- 6.99
Iteration 30/100: rewards 18.08 +/- 7.67
Iteration 31/100: rewards 19.64 +/- 8.65
Iteration 32/100: rewards 19.96 +/- 8.52
Iteration 33/100: rewards 20.22 +/- 10.6
Iteration 34/100: rewards 22.32 +/- 11.55
Iteration 35/100: rewards 17.38 +/- 8.66
Iteration 36/100: rewards 19.14 +/- 9.56
Iteration 37/100: rewards 20.48 +/- 8.96
Iteration 38/100: rewards 18.16 +/- 7.62
Iteration 39/100: rewards 18.62 +/- 6.66
Iteration 40/100: rewards 20.12 +/- 9.39
Iteration 41/100: rewards 19.96 +/- 9.97
Iteration 42/100: rewards 17.58 +/- 6.65
Iteration 43/100: rewards 20.04 +/- 10.36
Iteration 44/100: rewards 17.6 +/- 7.32
Iteration 45/100: rewards 19.64 +/- 10.05
Iteration 46/100: rewards 19.34 +/- 12.17
Iteration 47/100: rewards 17.76 +/- 7.42
Iteration 48/100: rewards 20.24 +/- 10.68
Iteration 49/100: rewards 21.38 +/- 9.21
Iteration 50/100: rewards 19.02 +/- 8.99
Iteration 51/100: rewards 18.74 +/- 7.92
Iteration 52/100: rewards 18.62 +/- 8.55
Iteration 53/100: rewards 19.56 +/- 12.44
Iteration 54/100: rewards 19.2 +/- 9.67
Iteration 55/100: rewards 17.36 +/- 7.52
Iteration 56/100: rewards 17.82 +/- 7.41
Iteration 57/100: rewards 19.9 +/- 9.6
Iteration 58/100: rewards 18.78 +/- 6.42
Iteration 59/100: rewards 20.04 +/- 10.56
Iteration 60/100: rewards 18.2 +/- 11.72
Iteration 61/100: rewards 18.78 +/- 7.84
Iteration 62/100: rewards 19.3 +/- 7.96
Iteration 63/100: rewards 19.5 +/- 9.41
Iteration 64/100: rewards 21.1 +/- 10.87
Iteration 65/100: rewards 18.38 +/- 7.41
Iteration 66/100: rewards 22.02 +/- 11.24
Iteration 67/100: rewards 18.8 +/- 7.57
Iteration 68/100: rewards 18.42 +/- 8.22
Iteration 69/100: rewards 19.04 +/- 8.95
Iteration 70/100: rewards 18.96 +/- 8.65
Iteration 71/100: rewards 18.64 +/- 9.11
Iteration 72/100: rewards 18.02 +/- 7.32
Iteration 73/100: rewards 18.14 +/- 9.15
Iteration 74/100: rewards 21.82 +/- 11.3
Iteration 75/100: rewards 16.9 +/- 6.28
Iteration 76/100: rewards 20.24 +/- 9.12
Iteration 77/100: rewards 19.3 +/- 10.26
Iteration 78/100: rewards 22.2 +/- 12.55
Iteration 79/100: rewards 22.1 +/- 12.78
Iteration 80/100: rewards 20.86 +/- 9.05
Iteration 81/100: rewards 22.4 +/- 12.89
Iteration 82/100: rewards 19.04 +/- 8.1
Iteration 83/100: rewards 20.24 +/- 9.84
Iteration 84/100: rewards 20.12 +/- 12.66
Iteration 85/100: rewards 20.6 +/- 9.98
Iteration 86/100: rewards 20.72 +/- 12.57
Iteration 87/100: rewards 21.22 +/- 8.9
Iteration 88/100: rewards 20.18 +/- 9.07
Iteration 89/100: rewards 21.46 +/- 10.15
Iteration 90/100: rewards 21.0 +/- 10.62
Iteration 91/100: rewards 20.4 +/- 10.48
Iteration 92/100: rewards 19.08 +/- 7.97
Iteration 93/100: rewards 19.98 +/- 9.12
Iteration 94/100: rewards 17.76 +/- 6.2
Iteration 95/100: rewards 20.18 +/- 9.61
Iteration 96/100: rewards 20.26 +/- 9.45
Iteration 97/100: rewards 18.98 +/- 9.21
Iteration 98/100: rewards 19.9 +/- 9.15
Iteration 99/100: rewards 22.3 +/- 9.43
Iteration 100/100: rewards 19.16 +/- 10.21
Iteration 1/100: rewards 17.8 +/- 6.91
Iteration 2/100: rewards 19.86 +/- 9.96
Iteration 3/100: rewards 17.78 +/- 7.26
Iteration 4/100: rewards 19.92 +/- 14.35
Iteration 5/100: rewards 20.82 +/- 9.36
Iteration 6/100: rewards 22.1 +/- 12.22
Iteration 7/100: rewards 25.1 +/- 15.89
Iteration 8/100: rewards 22.36 +/- 10.46
Iteration 9/100: rewards 26.64 +/- 15.97
Iteration 10/100: rewards 24.9 +/- 10.85
Iteration 11/100: rewards 27.4 +/- 15.11
Iteration 12/100: rewards 25.36 +/- 14.72
Iteration 13/100: rewards 23.0 +/- 12.22
Iteration 14/100: rewards 28.2 +/- 13.55
Iteration 15/100: rewards 27.3 +/- 15.19
Iteration 16/100: rewards 28.12 +/- 16.14
Iteration 17/100: rewards 30.34 +/- 15.64
Iteration 18/100: rewards 27.88 +/- 14.27
Iteration 19/100: rewards 29.76 +/- 14.35
Iteration 20/100: rewards 31.36 +/- 16.38
Iteration 21/100: rewards 39.6 +/- 19.74
Iteration 22/100: rewards 30.32 +/- 13.43
Iteration 23/100: rewards 33.56 +/- 15.85
Iteration 24/100: rewards 39.12 +/- 23.75
Iteration 25/100: rewards 37.96 +/- 19.38
Iteration 26/100: rewards 42.56 +/- 20.74
Iteration 27/100: rewards 45.62 +/- 20.8
Iteration 28/100: rewards 46.22 +/- 22.9
Iteration 29/100: rewards 41.3 +/- 22.02
Iteration 30/100: rewards 51.04 +/- 23.52
Iteration 31/100: rewards 45.86 +/- 21.79
Iteration 32/100: rewards 44.4 +/- 25.26
Iteration 33/100: rewards 54.9 +/- 31.43
Iteration 34/100: rewards 52.18 +/- 21.58
Iteration 35/100: rewards 56.26 +/- 24.99
Iteration 36/100: rewards 63.9 +/- 31.21
Iteration 37/100: rewards 57.64 +/- 29.58
Iteration 38/100: rewards 70.42 +/- 28.01
Iteration 39/100: rewards 73.48 +/- 31.92
Iteration 40/100: rewards 72.18 +/- 30.24
Iteration 41/100: rewards 66.52 +/- 22.44
Iteration 42/100: rewards 70.14 +/- 27.35
Iteration 43/100: rewards 73.68 +/- 31.13
Iteration 44/100: rewards 75.88 +/- 30.86
Iteration 45/100: rewards 83.02 +/- 32.75
Iteration 46/100: rewards 76.8 +/- 24.77
Iteration 47/100: rewards 83.3 +/- 29.31
Iteration 48/100: rewards 89.76 +/- 37.4
Iteration 49/100: rewards 94.0 +/- 46.38
Iteration 50/100: rewards 80.2 +/- 33.76
Iteration 51/100: rewards 85.18 +/- 36.12
Iteration 52/100: rewards 79.26 +/- 25.86
Iteration 53/100: rewards 81.94 +/- 31.21
Iteration 54/100: rewards 90.16 +/- 36.5
Iteration 55/100: rewards 93.44 +/- 33.35
Iteration 56/100: rewards 108.22 +/- 36.04
Iteration 57/100: rewards 129.68 +/- 49.8
Iteration 58/100: rewards 135.52 +/- 49.77
Iteration 59/100: rewards 146.02 +/- 53.63
Iteration 60/100: rewards 150.52 +/- 55.38
Iteration 61/100: rewards 162.5 +/- 59.57
Iteration 62/100: rewards 194.4 +/- 49.16
Iteration 63/100: rewards 186.08 +/- 51.91
Iteration 64/100: rewards 195.94 +/- 57.22
Iteration 65/100: rewards 198.44 +/- 61.17
Iteration 66/100: rewards 225.82 +/- 57.65
Iteration 67/100: rewards 252.76 +/- 71.23
Iteration 68/100: rewards 290.14 +/- 92.08
Iteration 69/100: rewards 334.46 +/- 101.44
Iteration 70/100: rewards 400.66 +/- 106.32
Iteration 71/100: rewards 435.62 +/- 96.94
Iteration 72/100: rewards 477.46 +/- 69.01
Iteration 73/100: rewards 487.4 +/- 48.55
Iteration 74/100: rewards 414.26 +/- 97.55
Iteration 75/100: rewards 357.56 +/- 88.73
Iteration 76/100: rewards 359.42 +/- 86.68
Iteration 77/100: rewards 380.9 +/- 88.71
Iteration 78/100: rewards 425.0 +/- 64.62
Iteration 79/100: rewards 459.66 +/- 61.95
Iteration 80/100: rewards 497.64 +/- 9.99
Iteration 81/100: rewards 494.9 +/- 18.03
Iteration 82/100: rewards 484.34 +/- 29.54
Iteration 83/100: rewards 446.94 +/- 56.17
Iteration 84/100: rewards 444.66 +/- 59.19
Iteration 85/100: rewards 489.18 +/- 26.03
Iteration 86/100: rewards 498.06 +/- 13.58
Iteration 87/100: rewards 496.42 +/- 13.25
Iteration 88/100: rewards 477.04 +/- 36.56
Iteration 89/100: rewards 449.94 +/- 52.12
Iteration 90/100: rewards 425.32 +/- 59.45
Iteration 91/100: rewards 462.52 +/- 39.03
Iteration 92/100: rewards 493.24 +/- 22.34
Iteration 93/100: rewards 500.0 +/- 0.0
Iteration 94/100: rewards 491.92 +/- 56.56
Iteration 95/100: rewards 496.28 +/- 14.86
Iteration 96/100: rewards 446.66 +/- 59.39
Iteration 97/100: rewards 453.72 +/- 58.74
Iteration 98/100: rewards 467.3 +/- 83.88
Iteration 99/100: rewards 484.08 +/- 78.02
Iteration 100/100: rewards 500.0 +/- 0.0
Iteration 1/100: rewards 17.8 +/- 6.91
Iteration 2/100: rewards 19.86 +/- 9.96
Iteration 3/100: rewards 17.78 +/- 7.26
Iteration 4/100: rewards 19.92 +/- 14.35
Iteration 5/100: rewards 20.82 +/- 9.36
Iteration 6/100: rewards 22.1 +/- 12.22
Iteration 7/100: rewards 25.1 +/- 15.89
Iteration 8/100: rewards 22.36 +/- 10.46
Iteration 9/100: rewards 27.14 +/- 14.57
Iteration 10/100: rewards 26.06 +/- 13.39
Iteration 11/100: rewards 23.62 +/- 9.82
Iteration 12/100: rewards 26.78 +/- 16.5
Iteration 13/100: rewards 30.2 +/- 17.55
Iteration 14/100: rewards 27.14 +/- 12.95
Iteration 15/100: rewards 26.12 +/- 15.29
Iteration 16/100: rewards 26.9 +/- 13.9
Iteration 17/100: rewards 27.34 +/- 14.59
Iteration 18/100: rewards 30.52 +/- 15.39
Iteration 19/100: rewards 28.94 +/- 14.54
Iteration 20/100: rewards 38.38 +/- 23.45
Iteration 21/100: rewards 35.88 +/- 22.93
Iteration 22/100: rewards 32.94 +/- 18.3
Iteration 23/100: rewards 38.24 +/- 22.43
Iteration 24/100: rewards 33.7 +/- 17.81
Iteration 25/100: rewards 37.8 +/- 21.25
Iteration 26/100: rewards 41.58 +/- 22.97
Iteration 27/100: rewards 40.06 +/- 23.02
Iteration 28/100: rewards 45.0 +/- 21.92
Iteration 29/100: rewards 40.92 +/- 20.72
Iteration 30/100: rewards 47.38 +/- 21.09
Iteration 31/100: rewards 49.66 +/- 27.89
Iteration 32/100: rewards 42.6 +/- 25.23
Iteration 33/100: rewards 55.0 +/- 32.11
Iteration 34/100: rewards 49.66 +/- 24.94
Iteration 35/100: rewards 53.48 +/- 24.56
Iteration 36/100: rewards 61.52 +/- 35.83
Iteration 37/100: rewards 53.24 +/- 25.66
Iteration 38/100: rewards 76.14 +/- 33.9
Iteration 39/100: rewards 63.1 +/- 24.88
Iteration 40/100: rewards 70.38 +/- 29.96
Iteration 41/100: rewards 84.22 +/- 36.4
Iteration 42/100: rewards 80.84 +/- 35.2
Iteration 43/100: rewards 85.66 +/- 31.83
Iteration 44/100: rewards 94.5 +/- 43.41
Iteration 45/100: rewards 100.74 +/- 34.97
Iteration 46/100: rewards 110.96 +/- 52.38
Iteration 47/100: rewards 111.78 +/- 43.77
Iteration 48/100: rewards 119.58 +/- 53.84
Iteration 49/100: rewards 136.34 +/- 55.69
Iteration 50/100: rewards 137.92 +/- 48.7
Iteration 51/100: rewards 163.72 +/- 53.39
Iteration 52/100: rewards 162.5 +/- 46.18
Iteration 53/100: rewards 170.1 +/- 54.5
Iteration 54/100: rewards 169.22 +/- 47.0
Iteration 55/100: rewards 190.56 +/- 52.74
Iteration 56/100: rewards 202.44 +/- 69.8
Iteration 57/100: rewards 189.28 +/- 53.36
Iteration 58/100: rewards 199.1 +/- 50.01
Iteration 59/100: rewards 226.96 +/- 80.0
Iteration 60/100: rewards 240.24 +/- 77.58
Iteration 61/100: rewards 238.04 +/- 94.83
Iteration 62/100: rewards 268.78 +/- 100.9
Iteration 63/100: rewards 330.4 +/- 126.38
Iteration 64/100: rewards 404.94 +/- 103.79
Iteration 65/100: rewards 387.04 +/- 129.64
Iteration 66/100: rewards 419.34 +/- 114.86
Iteration 67/100: rewards 369.44 +/- 116.67
Iteration 68/100: rewards 340.5 +/- 121.3
Iteration 69/100: rewards 401.92 +/- 95.32
Iteration 70/100: rewards 437.72 +/- 96.32
Iteration 71/100: rewards 483.68 +/- 61.79
Iteration 72/100: rewards 493.42 +/- 33.44
Iteration 73/100: rewards 500.0 +/- 0.0
Iteration 74/100: rewards 484.94 +/- 73.81
Iteration 75/100: rewards 492.06 +/- 55.58
Iteration 76/100: rewards 492.78 +/- 35.17
Iteration 77/100: rewards 495.34 +/- 32.62
Iteration 78/100: rewards 491.72 +/- 57.96
Iteration 79/100: rewards 481.68 +/- 87.62
Iteration 80/100: rewards 494.4 +/- 39.2
Iteration 81/100: rewards 500.0 +/- 0.0
Iteration 82/100: rewards 500.0 +/- 0.0
Iteration 83/100: rewards 485.84 +/- 73.67
Iteration 84/100: rewards 498.64 +/- 9.52
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 490.66 +/- 65.38
Iteration 90/100: rewards 490.54 +/- 66.22
Iteration 91/100: rewards 474.48 +/- 77.25
Iteration 92/100: rewards 211.18 +/- 39.26
Iteration 93/100: rewards 132.88 +/- 38.7
Iteration 94/100: rewards 97.94 +/- 38.95
Iteration 95/100: rewards 76.02 +/- 40.52
Iteration 96/100: rewards 62.84 +/- 38.94
Iteration 97/100: rewards 49.0 +/- 33.6
Iteration 98/100: rewards 43.16 +/- 30.93
Iteration 99/100: rewards 31.74 +/- 20.83
Iteration 100/100: rewards 27.76 +/- 17.89
Iteration 1/100: rewards 17.8 +/- 6.91
Iteration 2/100: rewards 18.44 +/- 8.44
Iteration 3/100: rewards 19.42 +/- 11.75
Iteration 4/100: rewards 21.34 +/- 9.61
Iteration 5/100: rewards 18.28 +/- 9.04
Iteration 6/100: rewards 20.28 +/- 11.91
Iteration 7/100: rewards 18.04 +/- 8.19
Iteration 8/100: rewards 21.48 +/- 11.12
Iteration 9/100: rewards 20.92 +/- 8.89
Iteration 10/100: rewards 20.28 +/- 10.97
Iteration 11/100: rewards 20.56 +/- 11.02
Iteration 12/100: rewards 20.38 +/- 11.69
Iteration 13/100: rewards 21.14 +/- 9.44
Iteration 14/100: rewards 21.02 +/- 13.52
Iteration 15/100: rewards 22.18 +/- 10.63
Iteration 16/100: rewards 20.72 +/- 10.17
Iteration 17/100: rewards 23.18 +/- 14.78
Iteration 18/100: rewards 20.4 +/- 8.41
Iteration 19/100: rewards 19.68 +/- 10.34
Iteration 20/100: rewards 20.02 +/- 9.43
Iteration 21/100: rewards 19.4 +/- 6.23
Iteration 22/100: rewards 22.2 +/- 11.16
Iteration 23/100: rewards 19.22 +/- 6.54
Iteration 24/100: rewards 21.72 +/- 9.24
Iteration 25/100: rewards 20.58 +/- 10.06
Iteration 26/100: rewards 22.4 +/- 12.13
Iteration 27/100: rewards 20.06 +/- 9.94
Iteration 28/100: rewards 22.62 +/- 10.67
Iteration 29/100: rewards 23.48 +/- 13.5
Iteration 30/100: rewards 22.86 +/- 14.98
Iteration 31/100: rewards 20.62 +/- 9.15
Iteration 32/100: rewards 25.38 +/- 11.63
Iteration 33/100: rewards 21.1 +/- 10.41
Iteration 34/100: rewards 22.48 +/- 9.69
Iteration 35/100: rewards 22.34 +/- 10.36
Iteration 36/100: rewards 25.42 +/- 14.33
Iteration 37/100: rewards 21.82 +/- 10.39
Iteration 38/100: rewards 21.54 +/- 12.19
Iteration 39/100: rewards 21.02 +/- 8.84
Iteration 40/100: rewards 21.6 +/- 9.41
Iteration 41/100: rewards 20.3 +/- 8.3
Iteration 42/100: rewards 22.94 +/- 13.43
Iteration 43/100: rewards 19.98 +/- 8.4
Iteration 44/100: rewards 24.02 +/- 10.69
Iteration 45/100: rewards 22.32 +/- 10.24
Iteration 46/100: rewards 21.24 +/- 7.74
Iteration 47/100: rewards 23.58 +/- 11.41
Iteration 48/100: rewards 23.68 +/- 13.82
Iteration 49/100: rewards 24.78 +/- 14.18
Iteration 50/100: rewards 23.58 +/- 13.07
Iteration 51/100: rewards 23.06 +/- 13.39
Iteration 52/100: rewards 23.68 +/- 13.02
Iteration 53/100: rewards 23.0 +/- 11.37
Iteration 54/100: rewards 22.9 +/- 10.63
Iteration 55/100: rewards 24.24 +/- 17.52
Iteration 56/100: rewards 24.64 +/- 14.23
Iteration 57/100: rewards 24.22 +/- 12.08
Iteration 58/100: rewards 27.36 +/- 18.29
Iteration 59/100: rewards 23.62 +/- 12.47
Iteration 60/100: rewards 24.06 +/- 10.43
Iteration 61/100: rewards 29.14 +/- 16.05
Iteration 62/100: rewards 25.24 +/- 11.83
Iteration 63/100: rewards 24.82 +/- 12.04
Iteration 64/100: rewards 29.44 +/- 19.21
Iteration 65/100: rewards 23.52 +/- 12.64
Iteration 66/100: rewards 25.72 +/- 17.24
Iteration 67/100: rewards 25.24 +/- 12.92
Iteration 68/100: rewards 26.6 +/- 18.86
Iteration 69/100: rewards 28.56 +/- 15.6
Iteration 70/100: rewards 26.68 +/- 13.7
Iteration 71/100: rewards 24.68 +/- 11.64
Iteration 72/100: rewards 22.4 +/- 10.2
Iteration 73/100: rewards 28.4 +/- 24.09
Iteration 74/100: rewards 29.26 +/- 17.22
Iteration 75/100: rewards 24.12 +/- 13.37
Iteration 76/100: rewards 25.34 +/- 11.5
Iteration 77/100: rewards 25.74 +/- 14.01
Iteration 78/100: rewards 25.94 +/- 13.64
Iteration 79/100: rewards 25.66 +/- 13.59
Iteration 80/100: rewards 28.3 +/- 13.78
Iteration 81/100: rewards 31.32 +/- 17.66
Iteration 82/100: rewards 26.08 +/- 13.24
Iteration 83/100: rewards 27.24 +/- 16.9
Iteration 84/100: rewards 22.34 +/- 8.97
Iteration 85/100: rewards 25.34 +/- 12.12
Iteration 86/100: rewards 31.56 +/- 21.95
Iteration 87/100: rewards 28.26 +/- 16.03
Iteration 88/100: rewards 30.72 +/- 19.0
Iteration 89/100: rewards 28.28 +/- 17.18
Iteration 90/100: rewards 30.24 +/- 16.77
Iteration 91/100: rewards 30.4 +/- 18.52
Iteration 92/100: rewards 31.0 +/- 21.8
Iteration 93/100: rewards 28.92 +/- 15.9
Iteration 94/100: rewards 29.6 +/- 16.61
Iteration 95/100: rewards 29.66 +/- 17.79
Iteration 96/100: rewards 30.26 +/- 16.43
Iteration 97/100: rewards 31.2 +/- 18.45
Iteration 98/100: rewards 30.7 +/- 19.95
Iteration 99/100: rewards 32.72 +/- 14.49
Iteration 100/100: rewards 33.14 +/- 17.85
Iteration 1/100: rewards 17.8 +/- 6.91
Iteration 2/100: rewards 18.44 +/- 8.44
Iteration 3/100: rewards 19.42 +/- 11.75
Iteration 4/100: rewards 21.34 +/- 9.61
Iteration 5/100: rewards 18.28 +/- 9.04
Iteration 6/100: rewards 20.28 +/- 11.91
Iteration 7/100: rewards 18.04 +/- 8.19
Iteration 8/100: rewards 21.48 +/- 11.12
Iteration 9/100: rewards 20.92 +/- 8.89
Iteration 10/100: rewards 20.28 +/- 10.97
Iteration 11/100: rewards 20.56 +/- 11.02
Iteration 12/100: rewards 20.38 +/- 11.69
Iteration 13/100: rewards 21.14 +/- 9.44
Iteration 14/100: rewards 21.02 +/- 13.52
Iteration 15/100: rewards 22.18 +/- 10.63
Iteration 16/100: rewards 20.72 +/- 10.17
Iteration 17/100: rewards 23.18 +/- 14.78
Iteration 18/100: rewards 20.4 +/- 8.41
Iteration 19/100: rewards 19.68 +/- 10.34
Iteration 20/100: rewards 20.02 +/- 9.43
Iteration 21/100: rewards 19.4 +/- 6.23
Iteration 22/100: rewards 22.2 +/- 11.16
Iteration 23/100: rewards 19.22 +/- 6.54
Iteration 24/100: rewards 21.72 +/- 9.24
Iteration 25/100: rewards 20.58 +/- 10.06
Iteration 26/100: rewards 22.4 +/- 12.13
Iteration 27/100: rewards 20.06 +/- 9.94
Iteration 28/100: rewards 22.62 +/- 10.67
Iteration 29/100: rewards 23.48 +/- 13.5
Iteration 30/100: rewards 22.86 +/- 14.98
Iteration 31/100: rewards 20.62 +/- 9.15
Iteration 32/100: rewards 25.38 +/- 11.63
Iteration 33/100: rewards 21.1 +/- 10.41
Iteration 34/100: rewards 22.48 +/- 9.69
Iteration 35/100: rewards 22.34 +/- 10.36
Iteration 36/100: rewards 25.42 +/- 14.33
Iteration 37/100: rewards 21.82 +/- 10.39
Iteration 38/100: rewards 21.54 +/- 12.19
Iteration 39/100: rewards 21.54 +/- 10.4
Iteration 40/100: rewards 22.0 +/- 10.49
Iteration 41/100: rewards 22.84 +/- 10.52
Iteration 42/100: rewards 29.58 +/- 14.88
Iteration 43/100: rewards 22.02 +/- 11.4
Iteration 44/100: rewards 23.34 +/- 9.76
Iteration 45/100: rewards 25.08 +/- 14.79
Iteration 46/100: rewards 23.0 +/- 11.3
Iteration 47/100: rewards 23.1 +/- 12.79
Iteration 48/100: rewards 24.62 +/- 11.96
Iteration 49/100: rewards 29.7 +/- 15.64
Iteration 50/100: rewards 23.98 +/- 14.83
Iteration 51/100: rewards 21.9 +/- 11.24
Iteration 52/100: rewards 23.42 +/- 9.71
Iteration 53/100: rewards 22.68 +/- 10.55
Iteration 54/100: rewards 25.9 +/- 14.35
Iteration 55/100: rewards 23.16 +/- 13.25
Iteration 56/100: rewards 20.52 +/- 10.17
Iteration 57/100: rewards 27.76 +/- 18.26
Iteration 58/100: rewards 26.42 +/- 12.35
Iteration 59/100: rewards 24.14 +/- 15.42
Iteration 60/100: rewards 23.3 +/- 13.42
Iteration 61/100: rewards 23.16 +/- 12.89
Iteration 62/100: rewards 24.12 +/- 12.82
Iteration 63/100: rewards 25.3 +/- 14.79
Iteration 64/100: rewards 24.72 +/- 15.07
Iteration 65/100: rewards 25.9 +/- 14.5
Iteration 66/100: rewards 21.38 +/- 9.9
Iteration 67/100: rewards 26.96 +/- 13.17
Iteration 68/100: rewards 23.18 +/- 10.99
Iteration 69/100: rewards 25.46 +/- 12.55
Iteration 70/100: rewards 24.6 +/- 11.49
Iteration 71/100: rewards 28.12 +/- 19.27
Iteration 72/100: rewards 28.02 +/- 15.47
Iteration 73/100: rewards 30.36 +/- 18.75
Iteration 74/100: rewards 28.88 +/- 13.06
Iteration 75/100: rewards 24.24 +/- 13.26
Iteration 76/100: rewards 25.88 +/- 13.93
Iteration 77/100: rewards 23.84 +/- 13.57
Iteration 78/100: rewards 25.82 +/- 12.17
Iteration 79/100: rewards 28.52 +/- 18.46
Iteration 80/100: rewards 30.18 +/- 15.17
Iteration 81/100: rewards 28.58 +/- 19.17
Iteration 82/100: rewards 26.12 +/- 12.74
Iteration 83/100: rewards 29.72 +/- 16.1
Iteration 84/100: rewards 31.34 +/- 20.39
Iteration 85/100: rewards 24.82 +/- 9.65
Iteration 86/100: rewards 28.06 +/- 18.99
Iteration 87/100: rewards 25.24 +/- 11.79
Iteration 88/100: rewards 25.66 +/- 14.32
Iteration 89/100: rewards 27.3 +/- 12.75
Iteration 90/100: rewards 31.56 +/- 16.39
Iteration 91/100: rewards 27.36 +/- 15.97
Iteration 92/100: rewards 28.62 +/- 15.38
Iteration 93/100: rewards 29.5 +/- 19.49
Iteration 94/100: rewards 28.6 +/- 16.05
Iteration 95/100: rewards 28.7 +/- 16.81
Iteration 96/100: rewards 33.06 +/- 14.84
Iteration 97/100: rewards 29.3 +/- 16.13
Iteration 98/100: rewards 31.36 +/- 15.92
Iteration 99/100: rewards 33.56 +/- 21.15
Iteration 100/100: rewards 27.96 +/- 15.11
Iteration 1/100: rewards 17.8 +/- 6.91
Iteration 2/100: rewards 18.56 +/- 9.3
Iteration 3/100: rewards 18.18 +/- 9.35
Iteration 4/100: rewards 18.18 +/- 7.64
Iteration 5/100: rewards 17.4 +/- 8.27
Iteration 6/100: rewards 18.2 +/- 8.56
Iteration 7/100: rewards 17.06 +/- 8.04
Iteration 8/100: rewards 19.04 +/- 9.26
Iteration 9/100: rewards 19.22 +/- 11.09
Iteration 10/100: rewards 20.34 +/- 10.64
Iteration 11/100: rewards 19.46 +/- 9.32
Iteration 12/100: rewards 20.84 +/- 11.27
Iteration 13/100: rewards 20.52 +/- 11.57
Iteration 14/100: rewards 20.54 +/- 11.54
Iteration 15/100: rewards 20.82 +/- 11.28
Iteration 16/100: rewards 19.52 +/- 11.1
Iteration 17/100: rewards 18.8 +/- 9.53
Iteration 18/100: rewards 21.48 +/- 10.77
Iteration 19/100: rewards 21.5 +/- 10.5
Iteration 20/100: rewards 18.0 +/- 8.7
Iteration 21/100: rewards 19.28 +/- 9.1
Iteration 22/100: rewards 20.58 +/- 16.19
Iteration 23/100: rewards 19.14 +/- 7.75
Iteration 24/100: rewards 20.26 +/- 10.12
Iteration 25/100: rewards 17.7 +/- 6.38
Iteration 26/100: rewards 20.54 +/- 12.12
Iteration 27/100: rewards 18.24 +/- 8.09
Iteration 28/100: rewards 18.28 +/- 7.18
Iteration 29/100: rewards 18.28 +/- 6.99
Iteration 30/100: rewards 18.08 +/- 7.67
Iteration 31/100: rewards 19.64 +/- 8.65
Iteration 32/100: rewards 19.96 +/- 8.52
Iteration 33/100: rewards 20.22 +/- 10.6
Iteration 34/100: rewards 22.32 +/- 11.55
Iteration 35/100: rewards 17.38 +/- 8.66
Iteration 36/100: rewards 19.14 +/- 9.56
Iteration 37/100: rewards 20.48 +/- 8.96
Iteration 38/100: rewards 18.16 +/- 7.62
Iteration 39/100: rewards 18.62 +/- 6.66
Iteration 40/100: rewards 20.12 +/- 9.39
Iteration 41/100: rewards 19.96 +/- 9.97
Iteration 42/100: rewards 17.58 +/- 6.65
Iteration 43/100: rewards 20.04 +/- 10.36
Iteration 44/100: rewards 17.6 +/- 7.32
Iteration 45/100: rewards 19.64 +/- 10.05
Iteration 46/100: rewards 19.34 +/- 12.17
Iteration 47/100: rewards 17.76 +/- 7.42
Iteration 48/100: rewards 20.24 +/- 10.68
Iteration 49/100: rewards 21.38 +/- 9.21
Iteration 50/100: rewards 19.02 +/- 8.99
Iteration 51/100: rewards 18.74 +/- 7.92
Iteration 52/100: rewards 18.62 +/- 8.55
Iteration 53/100: rewards 19.56 +/- 12.44
Iteration 54/100: rewards 19.2 +/- 9.67
Iteration 55/100: rewards 17.36 +/- 7.52
Iteration 56/100: rewards 17.82 +/- 7.41
Iteration 57/100: rewards 19.9 +/- 9.6
Iteration 58/100: rewards 18.78 +/- 6.42
Iteration 59/100: rewards 20.04 +/- 10.56
Iteration 60/100: rewards 18.2 +/- 11.72
Iteration 61/100: rewards 18.78 +/- 7.84
Iteration 62/100: rewards 19.3 +/- 7.96
Iteration 63/100: rewards 19.5 +/- 9.41
Iteration 64/100: rewards 21.1 +/- 10.87
Iteration 65/100: rewards 18.38 +/- 7.41
Iteration 66/100: rewards 22.02 +/- 11.24
Iteration 67/100: rewards 18.8 +/- 7.57
Iteration 68/100: rewards 19.64 +/- 10.11
Iteration 69/100: rewards 19.34 +/- 9.24
Iteration 70/100: rewards 20.66 +/- 10.08
Iteration 71/100: rewards 19.54 +/- 10.48
Iteration 72/100: rewards 18.68 +/- 8.31
Iteration 73/100: rewards 22.34 +/- 12.04
Iteration 74/100: rewards 21.64 +/- 12.55
Iteration 75/100: rewards 20.76 +/- 8.99
Iteration 76/100: rewards 22.32 +/- 9.26
Iteration 77/100: rewards 22.24 +/- 11.98
Iteration 78/100: rewards 18.6 +/- 8.7
Iteration 79/100: rewards 20.6 +/- 10.68
Iteration 80/100: rewards 23.68 +/- 14.9
Iteration 81/100: rewards 20.56 +/- 11.95
Iteration 82/100: rewards 21.04 +/- 12.53
Iteration 83/100: rewards 18.36 +/- 8.19
Iteration 84/100: rewards 21.24 +/- 11.61
Iteration 85/100: rewards 21.72 +/- 13.2
Iteration 86/100: rewards 20.2 +/- 10.23
Iteration 87/100: rewards 22.34 +/- 12.27
Iteration 88/100: rewards 21.24 +/- 12.78
Iteration 89/100: rewards 19.26 +/- 9.38
Iteration 90/100: rewards 22.3 +/- 13.07
Iteration 91/100: rewards 19.12 +/- 9.15
Iteration 92/100: rewards 17.86 +/- 9.61
Iteration 93/100: rewards 18.6 +/- 7.42
Iteration 94/100: rewards 19.08 +/- 10.44
Iteration 95/100: rewards 20.04 +/- 10.42
Iteration 96/100: rewards 18.72 +/- 9.16
Iteration 97/100: rewards 18.56 +/- 8.79
Iteration 98/100: rewards 18.86 +/- 8.29
Iteration 99/100: rewards 21.04 +/- 10.97
Iteration 100/100: rewards 17.12 +/- 8.9
Iteration 1/100: rewards 17.8 +/- 6.91
Iteration 2/100: rewards 18.56 +/- 9.3
Iteration 3/100: rewards 18.18 +/- 9.35
Iteration 4/100: rewards 18.18 +/- 7.64
Iteration 5/100: rewards 17.4 +/- 8.27
Iteration 6/100: rewards 18.2 +/- 8.56
Iteration 7/100: rewards 17.06 +/- 8.04
Iteration 8/100: rewards 19.04 +/- 9.26
Iteration 9/100: rewards 19.22 +/- 11.09
Iteration 10/100: rewards 20.34 +/- 10.64
Iteration 11/100: rewards 19.46 +/- 9.32
Iteration 12/100: rewards 20.84 +/- 11.27
Iteration 13/100: rewards 20.52 +/- 11.57
Iteration 14/100: rewards 20.54 +/- 11.54
Iteration 15/100: rewards 20.82 +/- 11.28
Iteration 16/100: rewards 19.52 +/- 11.1
Iteration 17/100: rewards 18.8 +/- 9.53
Iteration 18/100: rewards 21.48 +/- 10.77
Iteration 19/100: rewards 21.5 +/- 10.5
Iteration 20/100: rewards 18.0 +/- 8.7
Iteration 21/100: rewards 19.28 +/- 9.1
Iteration 22/100: rewards 20.58 +/- 16.19
Iteration 23/100: rewards 19.14 +/- 7.75
Iteration 24/100: rewards 20.26 +/- 10.12
Iteration 25/100: rewards 17.7 +/- 6.38
Iteration 26/100: rewards 20.54 +/- 12.12
Iteration 27/100: rewards 18.24 +/- 8.09
Iteration 28/100: rewards 18.28 +/- 7.18
Iteration 29/100: rewards 18.28 +/- 6.99
Iteration 30/100: rewards 18.08 +/- 7.67
Iteration 31/100: rewards 19.64 +/- 8.65
Iteration 32/100: rewards 19.96 +/- 8.52
Iteration 33/100: rewards 20.22 +/- 10.6
Iteration 34/100: rewards 22.32 +/- 11.55
Iteration 35/100: rewards 17.38 +/- 8.66
Iteration 36/100: rewards 19.14 +/- 9.56
Iteration 37/100: rewards 20.48 +/- 8.96
Iteration 38/100: rewards 18.16 +/- 7.62
Iteration 39/100: rewards 18.62 +/- 6.66
Iteration 40/100: rewards 20.12 +/- 9.39
Iteration 41/100: rewards 19.96 +/- 9.97
Iteration 42/100: rewards 17.58 +/- 6.65
Iteration 43/100: rewards 20.04 +/- 10.36
Iteration 44/100: rewards 17.6 +/- 7.32
Iteration 45/100: rewards 19.64 +/- 10.05
Iteration 46/100: rewards 19.34 +/- 12.17
Iteration 47/100: rewards 17.76 +/- 7.42
Iteration 48/100: rewards 20.24 +/- 10.68
Iteration 49/100: rewards 21.38 +/- 9.21
Iteration 50/100: rewards 19.02 +/- 8.99
Iteration 51/100: rewards 18.74 +/- 7.92
Iteration 52/100: rewards 18.62 +/- 8.55
Iteration 53/100: rewards 19.56 +/- 12.44
Iteration 54/100: rewards 19.2 +/- 9.67
Iteration 55/100: rewards 17.36 +/- 7.52
Iteration 56/100: rewards 17.82 +/- 7.41
Iteration 57/100: rewards 19.9 +/- 9.6
Iteration 58/100: rewards 18.78 +/- 6.42
Iteration 59/100: rewards 20.04 +/- 10.56
Iteration 60/100: rewards 18.2 +/- 11.72
Iteration 61/100: rewards 18.78 +/- 7.84
Iteration 62/100: rewards 19.3 +/- 7.96
Iteration 63/100: rewards 19.5 +/- 9.41
Iteration 64/100: rewards 21.1 +/- 10.87
Iteration 65/100: rewards 18.38 +/- 7.41
Iteration 66/100: rewards 22.02 +/- 11.24
Iteration 67/100: rewards 18.8 +/- 7.57
Iteration 68/100: rewards 18.42 +/- 8.22
Iteration 69/100: rewards 19.04 +/- 8.95
Iteration 70/100: rewards 18.96 +/- 8.65
Iteration 71/100: rewards 18.64 +/- 9.11
Iteration 72/100: rewards 18.02 +/- 7.32
Iteration 73/100: rewards 18.14 +/- 9.15
Iteration 74/100: rewards 21.82 +/- 11.3
Iteration 75/100: rewards 16.9 +/- 6.28
Iteration 76/100: rewards 20.24 +/- 9.12
Iteration 77/100: rewards 19.3 +/- 10.26
Iteration 78/100: rewards 22.2 +/- 12.55
Iteration 79/100: rewards 22.1 +/- 12.78
Iteration 80/100: rewards 20.86 +/- 9.05
Iteration 81/100: rewards 22.4 +/- 12.89
Iteration 82/100: rewards 19.04 +/- 8.1
Iteration 83/100: rewards 20.24 +/- 9.84
Iteration 84/100: rewards 20.12 +/- 12.66
Iteration 85/100: rewards 20.6 +/- 9.98
Iteration 86/100: rewards 20.72 +/- 12.57
Iteration 87/100: rewards 21.22 +/- 8.9
Iteration 88/100: rewards 20.18 +/- 9.07
Iteration 89/100: rewards 21.46 +/- 10.15
Iteration 90/100: rewards 21.0 +/- 10.62
Iteration 91/100: rewards 20.4 +/- 10.48
Iteration 92/100: rewards 19.08 +/- 7.97
Iteration 93/100: rewards 19.98 +/- 9.12
Iteration 94/100: rewards 17.76 +/- 6.2
Iteration 95/100: rewards 20.68 +/- 10.21
Iteration 96/100: rewards 19.76 +/- 9.69
Iteration 97/100: rewards 18.98 +/- 9.21
Iteration 98/100: rewards 19.9 +/- 9.15
Iteration 99/100: rewards 22.3 +/- 9.43
Iteration 100/100: rewards 19.16 +/- 10.21
```

Out[18]:

```
<matplotlib.legend.Legend at 0x7f688beea090>
```

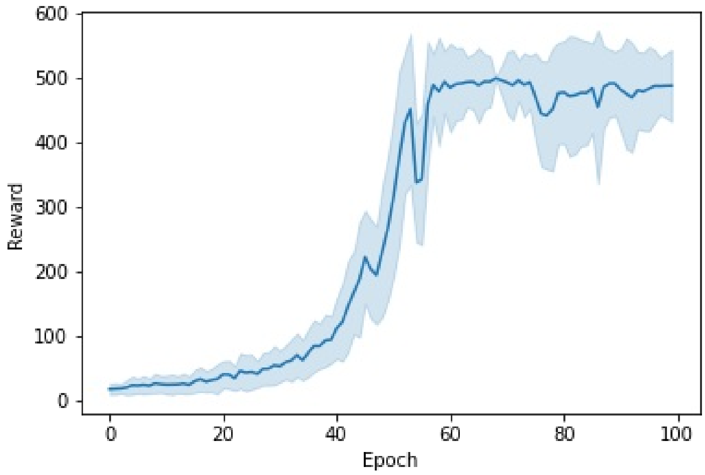


In [19]:

```
#0.99, 1e-2, 1e-2
# Provide your best config here and run this cell
config = {
    'env_id': 'CartPole-v1',
    'seed': 8953,
    'gamma': 0.99,
    'policy_layers': [16, 8],
    'policy_learning_rate': 1e-2,
    'use_baseline': True,
    'value_layers': [16, 8, 8],
    'value_learning_rate': 1e-2,
}
```

```
agent = ActorCriticAgent(config)
ActorCritic_rewards2 = agent.train(n_episodes=100, n_iterations=100)
```

```
Iteration 1/100: rewards 18.18 +/- 8.2
Iteration 2/100: rewards 18.72 +/- 8.63
Iteration 3/100: rewards 19.2 +/- 7.08
Iteration 4/100: rewards 20.69 +/- 11.44
Iteration 5/100: rewards 23.89 +/- 13.58
Iteration 6/100: rewards 23.49 +/- 11.38
Iteration 7/100: rewards 24.57 +/- 14.02
Iteration 8/100: rewards 23.24 +/- 11.42
Iteration 9/100: rewards 27.12 +/- 14.87
Iteration 10/100: rewards 25.83 +/- 12.97
Iteration 11/100: rewards 24.99 +/- 14.18
Iteration 12/100: rewards 25.23 +/- 15.5
Iteration 13/100: rewards 25.46 +/- 12.85
Iteration 14/100: rewards 26.8 +/- 15.6
Iteration 15/100: rewards 24.59 +/- 13.29
Iteration 16/100: rewards 30.38 +/- 17.43
Iteration 17/100: rewards 33.69 +/- 18.84
Iteration 18/100: rewards 29.6 +/- 16.01
Iteration 19/100: rewards 32.11 +/- 18.3
Iteration 20/100: rewards 33.77 +/- 22.84
Iteration 21/100: rewards 40.99 +/- 20.91
Iteration 22/100: rewards 40.6 +/- 21.71
Iteration 23/100: rewards 34.85 +/- 18.65
Iteration 24/100: rewards 46.38 +/- 27.18
Iteration 25/100: rewards 43.42 +/- 27.88
Iteration 26/100: rewards 44.69 +/- 26.65
Iteration 27/100: rewards 41.42 +/- 20.59
Iteration 28/100: rewards 49.44 +/- 25.17
Iteration 29/100: rewards 49.81 +/- 25.47
Iteration 30/100: rewards 54.96 +/- 29.8
Iteration 31/100: rewards 53.42 +/- 24.4
Iteration 32/100: rewards 59.9 +/- 26.05
Iteration 33/100: rewards 62.42 +/- 32.44
Iteration 34/100: rewards 70.93 +/- 33.71
Iteration 35/100: rewards 62.84 +/- 30.99
Iteration 36/100: rewards 74.55 +/- 37.16
Iteration 37/100: rewards 84.92 +/- 39.6
Iteration 38/100: rewards 84.75 +/- 34.91
Iteration 39/100: rewards 93.43 +/- 40.38
Iteration 40/100: rewards 94.62 +/- 36.29
Iteration 41/100: rewards 112.71 +/- 46.69
Iteration 42/100: rewards 121.83 +/- 60.06
Iteration 43/100: rewards 146.66 +/- 69.91
Iteration 44/100: rewards 167.5 +/- 64.21
Iteration 45/100: rewards 188.15 +/- 89.17
Iteration 46/100: rewards 222.58 +/- 71.77
Iteration 47/100: rewards 204.12 +/- 76.61
Iteration 48/100: rewards 194.61 +/- 75.15
Iteration 49/100: rewards 230.87 +/- 100.72
Iteration 50/100: rewards 267.29 +/- 111.51
Iteration 51/100: rewards 315.05 +/- 122.66
Iteration 52/100: rewards 375.44 +/- 134.77
Iteration 53/100: rewards 430.67 +/- 109.45
Iteration 54/100: rewards 452.06 +/- 116.62
Iteration 55/100: rewards 338.22 +/- 91.42
Iteration 56/100: rewards 343.15 +/- 101.28
Iteration 57/100: rewards 458.91 +/- 97.5
Iteration 58/100: rewards 489.45 +/- 48.36
Iteration 59/100: rewards 479.12 +/- 83.83
Iteration 60/100: rewards 494.33 +/- 47.5
Iteration 61/100: rewards 485.0 +/- 68.45
Iteration 62/100: rewards 490.8 +/- 55.94
Iteration 63/100: rewards 491.85 +/- 55.54
Iteration 64/100: rewards 493.89 +/- 38.42
Iteration 65/100: rewards 494.36 +/- 43.09
Iteration 66/100: rewards 488.99 +/- 57.77
Iteration 67/100: rewards 494.59 +/- 42.44
Iteration 68/100: rewards 494.47 +/- 39.11
Iteration 69/100: rewards 499.75 +/- 2.49
Iteration 70/100: rewards 496.61 +/- 22.34
Iteration 71/100: rewards 493.25 +/- 47.27
Iteration 72/100: rewards 488.95 +/- 54.5
Iteration 73/100: rewards 496.78 +/- 31.34
Iteration 74/100: rewards 489.57 +/- 49.41
Iteration 75/100: rewards 493.49 +/- 40.51
Iteration 76/100: rewards 470.93 +/- 67.99
Iteration 77/100: rewards 444.73 +/- 81.8
Iteration 78/100: rewards 442.08 +/- 82.74
Iteration 79/100: rewards 451.75 +/- 94.77
Iteration 80/100: rewards 476.71 +/- 78.31
Iteration 81/100: rewards 477.85 +/- 79.01
Iteration 82/100: rewards 472.22 +/- 93.23
Iteration 83/100: rewards 473.67 +/- 90.31
Iteration 84/100: rewards 477.15 +/- 82.2
Iteration 85/100: rewards 476.99 +/- 80.05
Iteration 86/100: rewards 484.45 +/- 69.36
Iteration 87/100: rewards 455.33 +/- 118.09
Iteration 88/100: rewards 486.04 +/- 63.62
Iteration 89/100: rewards 491.61 +/- 52.65
Iteration 90/100: rewards 491.71 +/- 49.43
Iteration 91/100: rewards 482.04 +/- 63.4
Iteration 92/100: rewards 476.04 +/- 85.52
Iteration 93/100: rewards 469.92 +/- 84.26
Iteration 94/100: rewards 481.18 +/- 59.61
Iteration 95/100: rewards 479.13 +/- 60.11
Iteration 96/100: rewards 483.44 +/- 65.25
Iteration 97/100: rewards 487.78 +/- 55.11
Iteration 98/100: rewards 487.55 +/- 43.54
Iteration 99/100: rewards 488.01 +/- 49.86
Iteration 100/100: rewards 488.3 +/- 55.25
```

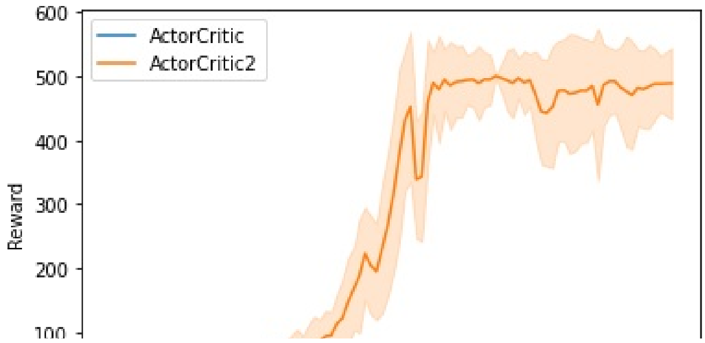


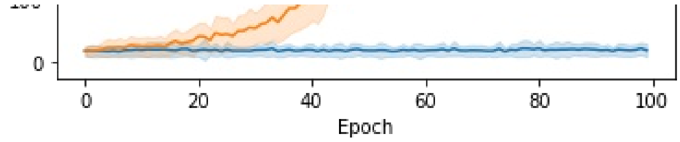
In [20]:

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

Out[20]:

```
<matplotlib.legend.Legend at 0x7f6889a52810>
```





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

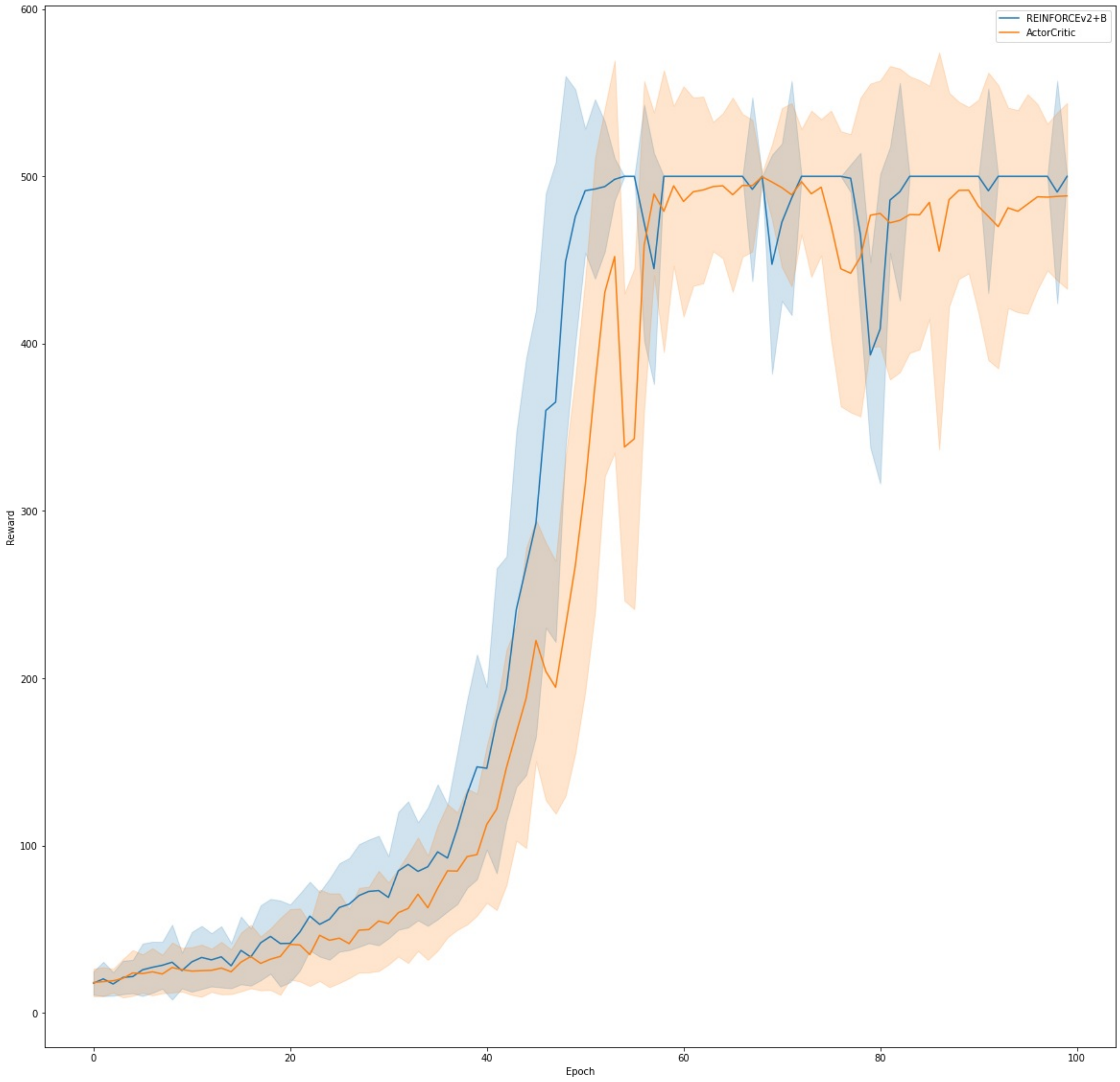
Report your observations and provide explanations for the same.

In [27]:

```
# 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'])
```

Out[27]:

```
<matplotlib.legend.Legend at 0x7f6898f8e150>
```



The actor critic method should have a lower variance

```
The actor-critic method should have a lower variance
and high biased compared to REINFORCE with baseline
so it is noisier. but here at the later episodes
the actor-critic has more variance, I think, cause of
the config the actor-critic did not train very
well in 100 episodes. also because of this problem
 the REINFORCE has better learning cure and it converged faster.
```

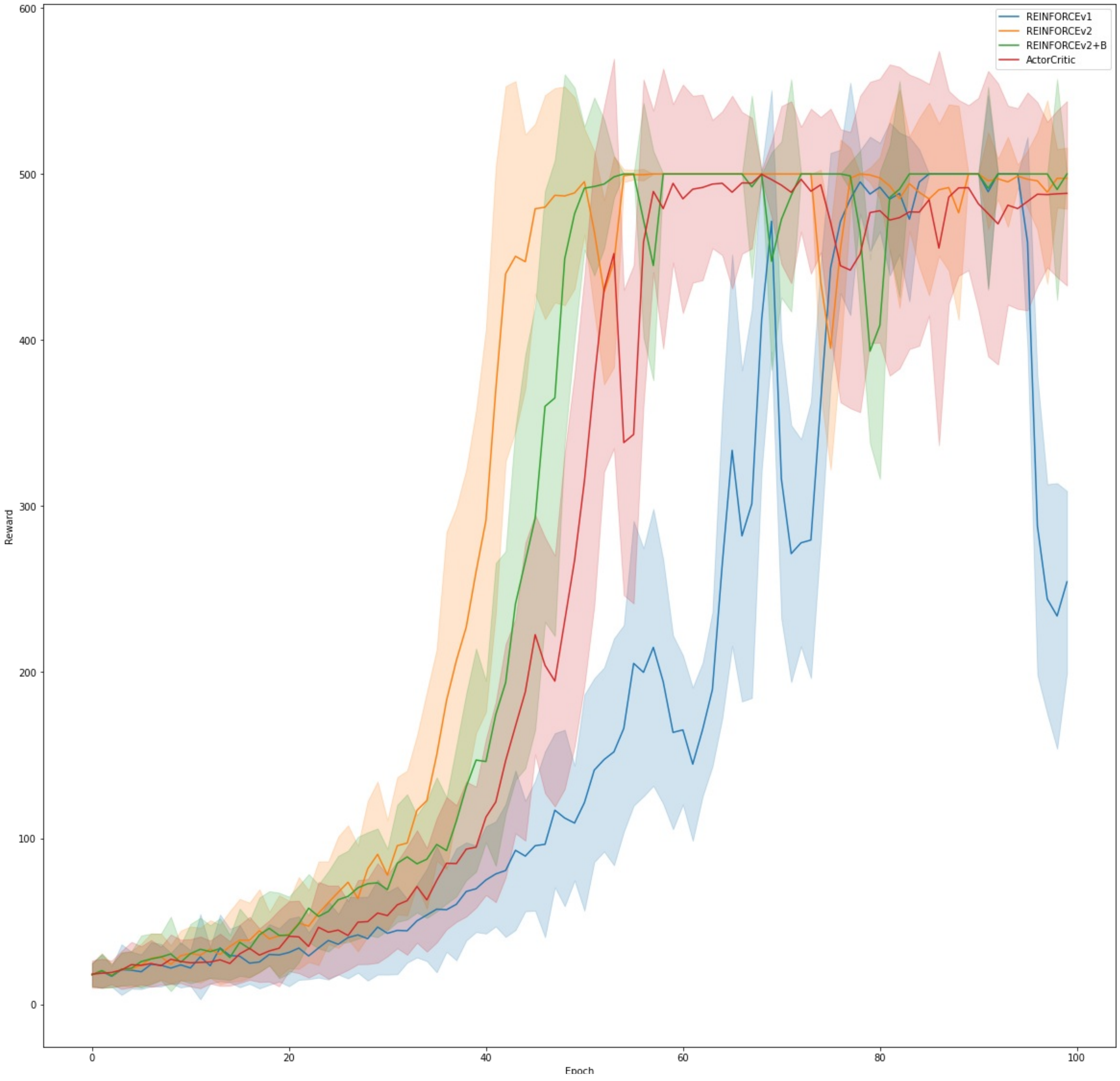
Qn 2.6: Plot all methods [2 Marks]

In [32]:

```
# 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'])
```

Out[32]:

```
<matplotlib.legend.Legend at 0x7f688979cc10>
```



Epoch