lr 10^-5 advs

May 3, 2023

[1]:
```
from learningBase import ReinforceAlgorithm
from environmentModelBase import Model, AdversaryModes
from neuralNetworkSimple import NNBase
import torch
import torch.nn as nn
from torch.distributions import Categorical
import numpy as np
import matplotlib.pyplot as plt
```

[ ]:
```
```

[2]:
```
hyperParams=[0.00001, 1, 0]
codeParams=[1, 10000, 1, 1]
```

[ ]:
```
for adv in range(len(AdversaryModes)):
    adversaryProbs=torch.zeros(len(AdversaryModes))
    adversaryProbs[adv]=1
    game = Model(totalDemand = 400,
                  tupleCosts = (57, 71),
                 totalStages = 25, adversaryProbs=adversaryProbs,␣
 ↪advHistoryNum=0)
    neuralNet=NNBase(num_input=game.T+2+game.advHistoryNum,␣
 ↪lr=hyperParams[0],num_actions=50)
    algorithm = ReinforceAlgorithm(game, neuralNet, numberIterations=3,␣
 ↪numberEpisodes=3_000_000, discountFactor =hyperParams[1])


    algorithm.solver(print_step=100_000,options=codeParams,converge_break=True)
```

```
policy reset
----------------------------------------
iter  0  stage  24  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0,
0, 0, 2, 0,
        0])
loss=  tensor(-0., grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.5663,
0.5663, 0.5663, 0.5663, 0.5663, 0.5663, 0.5663, 0.5663, 0.5663, 0.5663,
        0.5663, 0.5663, 0.5663, 0.5663, 0.5663, 0.5663, 0.5663, 0.5663, 0.5663,
```

```
        0.5663, 0.5663, 0.5663, 0.5663, 0.5663, 0.5663, 0.5663]) return=
139847.7232627179
probs of actions:  tensor([0.8786, 0.8836, 0.8852, 0.9041, 0.9160, 0.8963,
0.8981, 0.8989, 0.9127,
        0.9010, 0.9015, 0.9021, 0.0518, 0.8964, 0.8851, 0.9235, 0.9054, 0.0415,
        0.9164, 0.9023, 0.9071, 0.9030, 0.0111, 0.9274, 0.9877],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5112, 0.5366, 0.5495, 0.5560, 0.5592, 0.5609, 0.5617,
0.5621, 0.5623,
        0.5624, 0.5624, 0.5625, 0.5624, 0.5662, 0.5644, 0.5634, 0.5630, 0.5626,
        0.5664, 0.5644, 0.5635, 0.5630, 0.5623, 0.5701, 0.5663])
finalReturns:  tensor([0.])
----------------------------------------
iter  0  stage  23  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([ 0,  0,  4,  8, 14,  0,  0,  1,  0, 22, 20,  0,  5,  5,  0,
0, 16,  0,
         0,  0,  1,  0,  0, 22,  0])
loss=  tensor(0.0097, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.1279,
1.1279, 1.1279, 1.1279, 1.1279, 1.1279, 1.1279, 1.1279, 1.1279,
        1.1279, 1.1279, 1.1279, 1.1279, 1.1279, 1.1279, 1.1279, 1.1279, 1.1279,
        1.1279, 1.1279, 1.1279, 1.1279, 1.1279, 1.1279, 0.5635]) return=
145876.19064216543
probs of actions:  tensor([0.4697, 0.5450, 0.0549, 0.0497, 0.0049, 0.4636,
0.4981, 0.0854, 0.5431,
        0.0618, 0.0067, 0.5533, 0.0268, 0.0303, 0.4079, 0.6228, 0.0022, 0.5265,
        0.5884, 0.4364, 0.0926, 0.5506, 0.5699, 0.7906, 0.9886],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5112, 0.5366, 0.5479, 0.5646, 0.5776, 0.6343, 0.5978,
0.5799, 0.5750,
        0.5203, 0.6114, 0.6865, 0.6204, 0.6092, 0.6062, 0.5841, 0.5477, 0.6298,
        0.5957, 0.5790, 0.5706, 0.5704, 0.5664, 0.5161, 0.6491])
finalReturns:  tensor([0.0372, 0.0856])
----------------------------------------
iter  0  stage  22  ep  78431   adversary:  AdversaryModes.myopic
  actions:  tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22, 22, 22,
        22, 22, 22,  0, 22, 22,  0])
loss=  tensor(0.0009, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.8150,
1.8150, 1.8150, 1.8150, 1.8150, 1.8150, 1.8150, 1.8150, 1.8150,
        1.8150, 1.8150, 1.8150, 1.8150, 1.8150, 1.8150, 1.8150, 1.8150, 1.8150,
        1.8150, 1.8150, 1.8150, 1.8150, 1.8150, 1.1670, 0.5625]) return=
167444.53665689877
probs of actions:  tensor([0.9807, 0.9663, 0.9871, 0.9885, 0.9861, 0.9875,
0.9866, 0.9924, 0.9865,
        0.9892, 0.9886, 0.9758, 0.9819, 0.9908, 0.9891, 0.9866, 0.9716, 0.9830,
        0.9787, 0.9912, 0.9762, 0.0128, 0.9990, 0.9996, 0.9954],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5718, 0.6302, 0.6604, 0.6757, 0.6834, 0.6873,
```

```
0.6893, 0.6902,
        0.6907, 0.6910, 0.6911, 0.6911, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912,
        0.6912, 0.6912, 0.6912, 0.7396, 0.5996, 0.6447, 0.7161])
finalReturns:  tensor([0.1454, 0.1938, 0.1536])
---------------------------------------
iter  0  stage  21  ep  21111   adversary:  AdversaryModes.myopic
  actions:  tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22, 22, 22,
        22, 22,  0, 22, 22, 22,  0])
loss=  tensor(0.0010, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.3420,
2.3420, 2.3420, 2.3420, 2.3420, 2.3420, 2.3420, 2.3420, 2.3420,
        2.3420, 2.3420, 2.3420, 2.3420, 2.3420, 2.3420, 2.3420, 2.3420, 2.3420,
        2.3420, 2.3420, 2.3420, 2.3420, 1.6940, 1.0895, 0.5270]) return=
167326.75936424857
probs of actions:  tensor([0.9970, 0.9941, 0.9981, 0.9980, 0.9979, 0.9981,
0.9981, 0.9989, 0.9977,
        0.9984, 0.9982, 0.9962, 0.9969, 0.9986, 0.9983, 0.9977, 0.9953, 0.9971,
        0.9968, 0.9987, 0.0032, 0.9990, 0.9999, 1.0000, 0.9967],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5718, 0.6302, 0.6604, 0.6757, 0.6834, 0.6873,
0.6893, 0.6902,
        0.6907, 0.6910, 0.6911, 0.6911, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912,
        0.6912, 0.6912, 0.7396, 0.5996, 0.6447, 0.6677, 0.7278])
finalReturns:  tensor([0.2978, 0.3462, 0.3061, 0.2008])
---------------------------------------
iter  0  stage  20  ep  788   adversary:  AdversaryModes.myopic
  actions:  tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22, 22, 22,
        22,  0, 22, 22, 22, 22,  0])
loss=  tensor(0.0018, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.8407,
2.8407, 2.8407, 2.8407, 2.8407, 2.8407, 2.8407, 2.8407, 2.8407,
        2.8407, 2.8407, 2.8407, 2.8407, 2.8407, 2.8407, 2.8407, 2.8407, 2.8407,
        2.8407, 2.8407, 2.8407, 2.1927, 1.5882, 1.0257, 0.4987]) return=
167267.7526299165
probs of actions:  tensor([9.9804e-01, 9.9601e-01, 9.9866e-01, 9.9866e-01,
9.9858e-01, 9.9872e-01,
        9.9872e-01, 9.9925e-01, 9.9842e-01, 9.9895e-01, 9.9880e-01, 9.9754e-01,
        9.9786e-01, 9.9907e-01, 9.9890e-01, 9.9846e-01, 9.9682e-01, 9.9812e-01,
        9.9790e-01, 5.2534e-04, 9.9901e-01, 9.9924e-01, 9.9993e-01, 9.9998e-01,
        9.9636e-01], grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5718, 0.6302, 0.6604, 0.6757, 0.6834, 0.6873,
0.6893, 0.6902,
        0.6907, 0.6910, 0.6911, 0.6911, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912,
        0.6912, 0.7396, 0.5996, 0.6447, 0.6677, 0.6794, 0.7337])
finalReturns:  tensor([0.4844, 0.5328, 0.4927, 0.3875, 0.2350])
---------------------------------------
iter  0  stage  19  ep  0   adversary:  AdversaryModes.myopic
  actions:  tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
```

```
22, 22, 22,
        22, 22, 22, 22, 22, 22,  0])
loss=  tensor(0.0029, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.4906,
3.4906, 3.4906, 3.4906, 3.4906, 3.4906, 3.4906, 3.4906, 3.4906,
        3.4906, 3.4906, 3.4906, 3.4906, 3.4906, 3.4906, 3.4906, 3.4906, 3.4906,
        3.4906, 3.4906, 2.7510, 2.1030, 1.5197, 0.9827, 0.4792]) return=
168576.33348198733
probs of actions:  tensor([0.9980, 0.9960, 0.9987, 0.9987, 0.9986, 0.9987,
0.9987, 0.9992, 0.9984,
        0.9989, 0.9988, 0.9975, 0.9979, 0.9991, 0.9989, 0.9985, 0.9968, 0.9981,
        0.9979, 0.9991, 0.9989, 0.9993, 0.9999, 1.0000, 0.9964],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5718, 0.6302, 0.6604, 0.6757, 0.6834, 0.6873,
0.6893, 0.6902,
        0.6907, 0.6910, 0.6911, 0.6911, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912,
        0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.7396])
finalReturns:  tensor([0.7050, 0.7534, 0.7102, 0.6023, 0.4481, 0.2604])
----------------------------------------
iter  0  stage  18  ep  349   adversary:  AdversaryModes.myopic
  actions:  tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22, 22,  0,
        22, 22, 22, 22, 22, 22,  0])
loss=  tensor(0.0040, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.7776,
3.7776, 3.7776, 3.7776, 3.7776, 3.7776, 3.7776, 3.7776, 3.7776,
        3.7776, 3.7776, 3.7776, 3.7776, 3.7776, 3.7776, 3.7776, 3.7776, 3.7776,
        3.7776, 3.1295, 2.5250, 1.9625, 1.4355, 0.9369, 0.4601]) return=
167223.4464147788
probs of actions:  tensor([0.9981, 0.9962, 0.9988, 0.9987, 0.9987, 0.9988,
0.9988, 0.9993, 0.9985,
        0.9990, 0.9989, 0.9976, 0.9980, 0.9991, 0.9990, 0.9985, 0.9970, 0.0014,
        0.9991, 0.9995, 0.9988, 0.9994, 0.9999, 1.0000, 0.9965],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5718, 0.6302, 0.6604, 0.6757, 0.6834, 0.6873,
0.6893, 0.6902,
        0.6907, 0.6910, 0.6911, 0.6911, 0.6912, 0.6912, 0.6912, 0.6912, 0.7396,
        0.5996, 0.6447, 0.6677, 0.6794, 0.6853, 0.6882, 0.7381])
finalReturns:  tensor([0.9256, 0.9740, 0.9338, 0.8286, 0.6761, 0.4895, 0.2780])
----------------------------------------
iter  0  stage  17  ep  219   adversary:  AdversaryModes.myopic
  actions:  tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22, 22, 22,
        22, 22, 22, 22, 22, 22,  0])
loss=  tensor(0.0051, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.3999,
4.3999, 4.3999, 4.3999, 4.3999, 4.3999, 4.3999, 4.3999, 4.3999,
        4.3999, 4.3999, 4.3999, 4.3999, 4.3999, 4.3999, 4.3999, 4.3999, 4.3999,
        3.6604, 3.0123, 2.4290, 1.8920, 1.3885, 0.9093, 0.4481]) return=
168576.33348198733
probs of actions:  tensor([0.9983, 0.9966, 0.9989, 0.9989, 0.9988, 0.9989,
```

```
0.9989, 0.9994, 0.9986,
        0.9991, 0.9990, 0.9979, 0.9982, 0.9992, 0.9991, 0.9987, 0.9973, 0.9990,
        0.9992, 0.9996, 0.9990, 0.9995, 1.0000, 1.0000, 0.9965],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5718, 0.6302, 0.6604, 0.6757, 0.6834, 0.6873,
0.6893, 0.6902,
        0.6907, 0.6910, 0.6911, 0.6911, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912,
        0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.7396])
finalReturns:  tensor([1.1780, 1.2264, 1.1833, 1.0754, 0.9212, 0.7335, 0.5215,
0.2915])
----------------------------------------
iter  0  stage  16  ep  5672   adversary:  AdversaryModes.myopic
  actions:  tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
0, 22, 22,
        22, 22, 22, 22, 22, 22,  0])
loss=  tensor(0.0037, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.6630,
4.6630, 4.6630, 4.6630, 4.6630, 4.6630, 4.6630, 4.6630, 4.6630,
        4.6630, 4.6630, 4.6630, 4.6630, 4.6630, 4.6630, 4.6630, 4.6630, 4.0150,
        3.4105, 2.8480, 2.3210, 1.8223, 1.3456, 0.8855, 0.4380]) return=
167212.36521898463
probs of actions:  tensor([9.9920e-01, 9.9831e-01, 9.9951e-01, 9.9950e-01,
9.9944e-01, 9.9955e-01,
        9.9955e-01, 9.9975e-01, 9.9934e-01, 9.9959e-01, 9.9954e-01, 9.9893e-01,
        9.9910e-01, 9.9968e-01, 9.9960e-01, 5.5382e-04, 9.9908e-01, 9.9982e-01,
        9.9983e-01, 9.9998e-01, 9.9956e-01, 9.9994e-01, 9.9999e-01, 1.0000e+00,
        9.9578e-01], grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5718, 0.6302, 0.6604, 0.6757, 0.6834, 0.6873,
0.6893, 0.6902,
        0.6907, 0.6910, 0.6911, 0.6911, 0.6912, 0.6912, 0.7396, 0.5996, 0.6447,
        0.6677, 0.6794, 0.6853, 0.6882, 0.6897, 0.6905, 0.7392])
finalReturns:  tensor([1.4214, 1.4698, 1.4296, 1.3244, 1.1720, 0.9853, 0.7739,
0.5442, 0.3012])
----------------------------------------
iter  0  stage  15  ep  0   adversary:  AdversaryModes.myopic
  actions:  tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22, 22, 22,
        22, 22, 22, 22, 22, 22,  0])
loss=  tensor(0.0053, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.2692,
5.2692, 5.2692, 5.2692, 5.2692, 5.2692, 5.2692, 5.2692, 5.2692,
        5.2692, 5.2692, 5.2692, 5.2692, 5.2692, 5.2692, 5.2692, 4.5296, 3.8816,
        3.2983, 2.7613, 2.2577, 1.7786, 1.3173, 0.8693, 0.4310]) return=
168576.33348198733
probs of actions:  tensor([0.9992, 0.9983, 0.9995, 0.9995, 0.9994, 0.9996,
0.9996, 0.9998, 0.9993,
        0.9996, 0.9995, 0.9989, 0.9991, 0.9997, 0.9996, 0.9994, 0.9990, 0.9998,
        0.9998, 1.0000, 0.9996, 0.9999, 1.0000, 1.0000, 0.9958],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5718, 0.6302, 0.6604, 0.6757, 0.6834, 0.6873,
```

```
0.6893, 0.6902,
        0.6907, 0.6910, 0.6911, 0.6911, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912,
        0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.7396])
finalReturns:  tensor([1.6912, 1.7396, 1.6964, 1.5885, 1.4343, 1.2467, 1.0346,
0.8047, 0.5615,
        0.3086])
----------------------------------------
iter  0  stage  14  ep  0   adversary:  AdversaryModes.myopic
  actions:  tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22, 22, 22,
        22, 22, 22, 22, 22, 22,  0])
loss=  tensor(0.0066, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.6948,
5.6948, 5.6948, 5.6948, 5.6948, 5.6948, 5.6948, 5.6948, 5.6948,
        5.6948, 5.6948, 5.6948, 5.6948, 5.6948, 5.6948, 4.9552, 4.3072, 3.7239,
        3.1869, 2.6834, 2.2042, 1.7430, 1.2949, 0.8566, 0.4256]) return=
168576.33348198733
probs of actions:  tensor([0.9992, 0.9983, 0.9995, 0.9995, 0.9994, 0.9996,
0.9996, 0.9998, 0.9993,
        0.9996, 0.9995, 0.9989, 0.9991, 0.9997, 0.9996, 0.9994, 0.9990, 0.9998,
        0.9998, 1.0000, 0.9996, 0.9999, 1.0000, 1.0000, 0.9958],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5718, 0.6302, 0.6604, 0.6757, 0.6834, 0.6873,
0.6893, 0.6902,
        0.6907, 0.6910, 0.6911, 0.6911, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912,
        0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.7396])
finalReturns:  tensor([1.9567, 2.0051, 1.9620, 1.8541, 1.6999, 1.5122, 1.3002,
1.0702, 0.8271,
        0.5742, 0.3140])
----------------------------------------
iter  0  stage  13  ep  0   adversary:  AdversaryModes.myopic
  actions:  tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22, 22, 22,
        22, 22, 22, 22, 22, 22,  0])
loss=  tensor(0.0078, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.1164,
6.1164, 6.1164, 6.1164, 6.1164, 6.1164, 6.1164, 6.1164, 6.1164,
        6.1164, 6.1164, 6.1164, 6.1164, 6.1164, 5.3768, 4.7288, 4.1455, 3.6085,
        3.1049, 2.6258, 2.1645, 1.7165, 1.2782, 0.8472, 0.4216]) return=
168576.33348198733
probs of actions:  tensor([0.9992, 0.9983, 0.9995, 0.9995, 0.9994, 0.9996,
0.9996, 0.9998, 0.9993,
        0.9996, 0.9995, 0.9989, 0.9991, 0.9997, 0.9996, 0.9994, 0.9990, 0.9998,
        0.9998, 1.0000, 0.9996, 0.9999, 1.0000, 1.0000, 0.9958],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5718, 0.6302, 0.6604, 0.6757, 0.6834, 0.6873,
0.6893, 0.6902,
        0.6907, 0.6910, 0.6911, 0.6911, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912,
        0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.7396])
finalReturns:  tensor([2.2264, 2.2748, 2.2316, 2.1237, 1.9695, 1.7819, 1.5698,
```

```
1.3399, 1.0967,
        0.8438, 0.5836, 0.3180])
---------------------------------------
iter  0  stage  12  ep  0   adversary:  AdversaryModes.myopic
  actions:  tensor([22, 22,  0, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22, 22, 22,
        22, 22, 22, 22, 22, 22,  0])
loss=  tensor(0.0105, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.5345,
6.5345, 6.5345, 6.5345, 6.5345, 6.5345, 6.5345, 6.5345, 6.5345,
        6.5345, 6.5345, 6.5345, 6.5345, 5.7952, 5.1473, 4.5640, 4.0270, 3.5235,
        3.0444, 2.5831, 2.1351, 1.6968, 1.2658, 0.8401, 0.4186]) return=
167235.25059974194
probs of actions:  tensor([9.9920e-01, 9.9831e-01, 3.7954e-04, 9.9957e-01,
9.9935e-01, 9.9952e-01,
        9.9954e-01, 9.9975e-01, 9.9933e-01, 9.9959e-01, 9.9954e-01, 9.9893e-01,
        9.9910e-01, 9.9968e-01, 9.9960e-01, 9.9939e-01, 9.9896e-01, 9.9985e-01,
        9.9984e-01, 9.9998e-01, 9.9957e-01, 9.9994e-01, 9.9999e-01, 1.0000e+00,
        9.9578e-01], grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5718, 0.6786, 0.5708, 0.6296, 0.6601, 0.6756,
0.6834, 0.6873,
        0.6892, 0.6902, 0.6907, 0.6910, 0.6911, 0.6911, 0.6912, 0.6912, 0.6912,
        0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.7396])
finalReturns:  tensor([2.4990, 2.5474, 2.5042, 2.3963, 2.2421, 2.0545, 1.8424,
1.6125, 1.3693,
        1.1164, 0.8562, 0.5907, 0.3210])
---------------------------------------
iter  0  stage  11  ep  13   adversary:  AdversaryModes.myopic
  actions:  tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22, 22, 22,
        22, 22, 22, 22, 22, 22,  0])
loss=  tensor(0.0135, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.9511,
6.9511, 6.9511, 6.9511, 6.9511, 6.9511, 6.9511, 6.9511, 6.9511,
        6.9511, 6.9511, 6.9511, 6.2116, 5.5636, 4.9804, 4.4434, 3.9398, 3.4607,
        2.9994, 2.5514, 2.1131, 1.6821, 1.2565, 0.8349, 0.4163]) return=
168576.33348198733
probs of actions:  tensor([0.9992, 0.9983, 0.9995, 0.9995, 0.9994, 0.9996,
0.9996, 0.9998, 0.9993,
        0.9996, 0.9996, 0.9990, 0.9992, 0.9997, 0.9996, 0.9994, 0.9990, 0.9998,
        0.9998, 1.0000, 0.9996, 0.9999, 1.0000, 1.0000, 0.9957],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5718, 0.6302, 0.6604, 0.6757, 0.6834, 0.6873,
0.6893, 0.6902,
        0.6907, 0.6910, 0.6911, 0.6911, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912,
        0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.7396])
finalReturns:  tensor([2.7739, 2.8223, 2.7791, 2.6712, 2.5170, 2.3294, 2.1173,
1.8874, 1.6442,
        1.3913, 1.1311, 0.8655, 0.5959, 0.3233])
---------------------------------------
```

```
iter  0  stage  10  ep  0   adversary:  AdversaryModes.myopic
  actions:  tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22, 22, 22,
        22, 22, 22, 22, 22, 22,  0])
loss=  tensor(0.0155, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.3655,
7.3655, 7.3655, 7.3655, 7.3655, 7.3655, 7.3655, 7.3655, 7.3655,
        7.3655, 7.3655, 6.6261, 5.9782, 5.3950, 4.8580, 4.3544, 3.8753, 3.4141,
        2.9660, 2.5277, 2.0967, 1.6711, 1.2495, 0.8310, 0.4146]) return=
168576.33348198733
probs of actions:  tensor([0.9992, 0.9983, 0.9995, 0.9995, 0.9994, 0.9996,
0.9996, 0.9998, 0.9993,
        0.9996, 0.9996, 0.9990, 0.9992, 0.9997, 0.9996, 0.9994, 0.9990, 0.9998,
        0.9998, 1.0000, 0.9996, 0.9999, 1.0000, 1.0000, 0.9957],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5718, 0.6302, 0.6604, 0.6757, 0.6834, 0.6873,
0.6893, 0.6902,
        0.6907, 0.6910, 0.6911, 0.6911, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912,
        0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.7396])
finalReturns:  tensor([3.0504, 3.0988, 3.0557, 2.9478, 2.7936, 2.6059, 2.3939,
2.1639, 1.9208,
        1.6679, 1.4077, 1.1421, 0.8725, 0.5998, 0.3250])
----------------------------------------
iter  0  stage  9  ep  0   adversary:  AdversaryModes.myopic
  actions:  tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22, 22, 22,
        22, 22, 22, 22, 22, 22,  0])
loss=  tensor(0.0175, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.7784,
7.7784, 7.7784, 7.7784, 7.7784, 7.7784, 7.7784, 7.7784, 7.7784,
        7.7784, 7.0393, 6.3915, 5.8083, 5.2713, 4.7678, 4.2887, 3.8274, 3.3794,
        2.9411, 2.5101, 2.0845, 1.6629, 1.2443, 0.8280, 0.4134]) return=
168576.33348198733
probs of actions:  tensor([0.9992, 0.9983, 0.9995, 0.9995, 0.9994, 0.9996,
0.9996, 0.9998, 0.9993,
        0.9996, 0.9996, 0.9990, 0.9992, 0.9997, 0.9996, 0.9994, 0.9990, 0.9998,
        0.9998, 1.0000, 0.9996, 0.9999, 1.0000, 1.0000, 0.9957],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5718, 0.6302, 0.6604, 0.6757, 0.6834, 0.6873,
0.6893, 0.6902,
        0.6907, 0.6910, 0.6911, 0.6911, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912,
        0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.7396])
finalReturns:  tensor([3.3282, 3.3766, 3.3335, 3.2256, 3.0714, 2.8838, 2.6717,
2.4418, 2.1986,
        1.9457, 1.6855, 1.4199, 1.1503, 0.8777, 0.6028, 0.3262])
----------------------------------------
iter  0  stage  8  ep  0   adversary:  AdversaryModes.myopic
  actions:  tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22, 22, 22,
        22, 22, 22, 22, 22, 22,  0])
```

```
loss=  tensor(0.0204, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.1899,
8.1899, 8.1899, 8.1899, 8.1899, 8.1899, 8.1899, 8.1899, 8.1899,
        7.4513, 6.8037, 6.2206, 5.6837, 5.1802, 4.7011, 4.2398, 3.7918, 3.3535,
        2.9225, 2.4969, 2.0753, 1.6567, 1.2404, 0.8258, 0.4124]) return=
168576.33348198733
probs of actions:  tensor([0.9992, 0.9983, 0.9995, 0.9995, 0.9994, 0.9996,
0.9996, 0.9998, 0.9993,
        0.9996, 0.9996, 0.9990, 0.9992, 0.9997, 0.9996, 0.9994, 0.9990, 0.9998,
        0.9998, 1.0000, 0.9996, 0.9999, 1.0000, 1.0000, 0.9957],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5718, 0.6302, 0.6604, 0.6757, 0.6834, 0.6873,
0.6893, 0.6902,
        0.6907, 0.6910, 0.6911, 0.6911, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912,
        0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.7396])
finalReturns:  tensor([3.6069, 3.6553, 3.6122, 3.5043, 3.3502, 3.1625, 2.9505,
2.7205, 2.4774,
        2.2245, 1.9643, 1.6987, 1.4291, 1.1565, 0.8816, 0.6050, 0.3272])
---------------------------------------
iter  0  stage  7  ep  0   adversary:  AdversaryModes.myopic
  actions:  tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22, 22, 22,
        22, 22, 22, 22, 22, 22,  0])
loss=  tensor(0.0220, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.5998,
8.5998, 8.5998, 8.5998, 8.5998, 8.5998, 8.5998, 7.8621,
        7.2150, 6.6321, 6.0953, 5.5919, 5.1128, 4.6515, 4.2035, 3.7652, 3.3342,
        2.9086, 2.4870, 2.0685, 1.6522, 1.2375, 0.8241, 0.4117]) return=
168576.33348198733
probs of actions:  tensor([0.9992, 0.9983, 0.9995, 0.9995, 0.9994, 0.9996,
0.9996, 0.9998, 0.9994,
        0.9996, 0.9996, 0.9990, 0.9992, 0.9997, 0.9996, 0.9995, 0.9990, 0.9998,
        0.9998, 1.0000, 0.9996, 0.9999, 1.0000, 1.0000, 0.9956],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5718, 0.6302, 0.6604, 0.6757, 0.6834, 0.6873,
0.6893, 0.6902,
        0.6907, 0.6910, 0.6911, 0.6911, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912,
        0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.7396])
finalReturns:  tensor([3.8863, 3.9347, 3.8916, 3.7838, 3.6296, 3.4420, 3.2300,
3.0000, 2.7569,
        2.5040, 2.2438, 1.9782, 1.7086, 1.4359, 1.1610, 0.8845, 0.6067, 0.3279])
---------------------------------------
iter  0  stage  6  ep  0   adversary:  AdversaryModes.myopic
  actions:  tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22, 22, 22,
        22, 22, 22, 22, 22, 22,  0])
loss=  tensor(0.0243, grad_fn=<NegBackward0>)   ,  base rewards= tensor([9.0073,
9.0073, 9.0073, 9.0073, 9.0073, 9.0073, 9.0073, 8.2716, 7.6254,
        7.0429, 6.5063, 6.0030, 5.5239, 5.0627, 4.6147, 4.1764, 3.7454, 3.3198,
        2.8982, 2.4797, 2.0633, 1.6487, 1.2353, 0.8229, 0.4112]) return=
```

```
168576.33348198733
probs of actions:  tensor([0.9992, 0.9983, 0.9995, 0.9995, 0.9994, 0.9996,
0.9996, 0.9998, 0.9994,
        0.9996, 0.9996, 0.9990, 0.9992, 0.9997, 0.9996, 0.9995, 0.9990, 0.9998,
        0.9998, 1.0000, 0.9996, 0.9999, 1.0000, 1.0000, 0.9956],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5718, 0.6302, 0.6604, 0.6757, 0.6834, 0.6873,
0.6893, 0.6902,
        0.6907, 0.6910, 0.6911, 0.6911, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912,
        0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.7396])
finalReturns:  tensor([4.1661, 4.2145, 4.1715, 4.0637, 3.9096, 3.7220, 3.5100,
3.2800, 3.0369,
        2.7840, 2.5238, 2.2582, 1.9886, 1.7159, 1.4411, 1.1645, 0.8867, 0.6079,
        0.3284])
----------------------------------------
iter  0  stage  5  ep  0   adversary:  AdversaryModes.myopic
  actions:  tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22, 22, 22,
        22, 22, 22, 22, 22, 22,  0])
loss=  tensor(0.0267, grad_fn=<NegBackward0>)   ,  base rewards= tensor([9.4107,
9.4107, 9.4107, 9.4107, 9.4107, 9.4107, 8.6789, 8.0345, 7.4529,
        6.9167, 6.4136, 5.9346, 5.4735, 5.0255, 4.5872, 4.1562, 3.7306, 3.3090,
        2.8904, 2.4741, 2.0595, 1.6461, 1.2337, 0.8220, 0.4108]) return=
168576.33348198733
probs of actions:  tensor([0.9992, 0.9983, 0.9995, 0.9995, 0.9994, 0.9996,
0.9996, 0.9998, 0.9994,
        0.9996, 0.9996, 0.9990, 0.9992, 0.9997, 0.9996, 0.9995, 0.9990, 0.9998,
        0.9998, 1.0000, 0.9996, 0.9999, 1.0000, 1.0000, 0.9956],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5718, 0.6302, 0.6604, 0.6757, 0.6834, 0.6873,
0.6893, 0.6902,
        0.6907, 0.6910, 0.6911, 0.6911, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912,
        0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.7396])
finalReturns:  tensor([4.4461, 4.4945, 4.4516, 4.3440, 4.1899, 4.0023, 3.7903,
3.5604, 3.3173,
        3.0644, 2.8042, 2.5386, 2.2690, 1.9964, 1.7215, 1.4449, 1.1671, 0.8883,
        0.6088, 0.3288])
----------------------------------------
iter  0  stage  4  ep  0   adversary:  AdversaryModes.myopic
  actions:  tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22, 22, 22,
        22, 22, 22, 22, 22, 22,  0])
loss=  tensor(0.0297, grad_fn=<NegBackward0>)   ,  base rewards= tensor([9.8065,
9.8065, 9.8065, 9.8065, 9.8065, 9.0824, 8.4416, 7.8618, 7.3264,
        6.8237, 6.3449, 5.8839, 5.4359, 4.9977, 4.5667, 4.1411, 3.7195, 3.3009,
        2.8846, 2.4700, 2.0566, 1.6442, 1.2325, 0.8213, 0.4105]) return=
168576.33348198733
probs of actions:  tensor([0.9992, 0.9983, 0.9995, 0.9995, 0.9994, 0.9996,
```

```
0.9996, 0.9998, 0.9994,
        0.9996, 0.9996, 0.9990, 0.9992, 0.9997, 0.9996, 0.9995, 0.9990, 0.9999,
        0.9998, 1.0000, 0.9996, 0.9999, 1.0000, 1.0000, 0.9956],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5718, 0.6302, 0.6604, 0.6757, 0.6834, 0.6873,
0.6893, 0.6902,
        0.6907, 0.6910, 0.6911, 0.6911, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912,
        0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.7396])
finalReturns:  tensor([4.7261, 4.7745, 4.7318, 4.6243, 4.4704, 4.2829, 4.0710,
3.8411, 3.5980,
        3.3451, 3.0849, 2.8193, 2.5497, 2.2771, 2.0022, 1.7256, 1.4478, 1.1690,
        0.8895, 0.6095, 0.3291])
----------------------------------------
iter  0  stage  3  ep  0   adversary:  AdversaryModes.myopic
  actions:  tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22, 22, 22,
        22, 22, 22, 22, 22, 22,  0])
loss=  tensor(0.0326, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([10.1875, 10.1875, 10.1875, 10.1875,  9.4787,  8.8452,  8.2688,  7.7351,
         7.2332,  6.7548,  6.2939,  5.8461,  5.4079,  4.9769,  4.5513,  4.1298,
         3.7112,  3.2949,  2.8803,  2.4669,  2.0545,  1.6427,  1.2316,  0.8208,
         0.4103]) return=  168576.33348198733
probs of actions:  tensor([0.9992, 0.9983, 0.9995, 0.9995, 0.9995, 0.9996,
0.9996, 0.9998, 0.9994,
        0.9996, 0.9996, 0.9991, 0.9992, 0.9997, 0.9996, 0.9995, 0.9990, 0.9999,
        0.9998, 1.0000, 0.9996, 0.9999, 1.0000, 1.0000, 0.9956],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5718, 0.6302, 0.6604, 0.6757, 0.6834, 0.6873,
0.6893, 0.6902,
        0.6907, 0.6910, 0.6911, 0.6911, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912,
        0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.7396])
finalReturns:  tensor([5.0054, 5.0538, 5.0116, 4.9046, 4.7510, 4.5637, 4.3518,
4.1220, 3.8789,
        3.6260, 3.3658, 3.1002, 2.8306, 2.5580, 2.2831, 2.0065, 1.7287, 1.4499,
        1.1705, 0.8904, 0.6100, 0.3293])
----------------------------------------
iter  0  stage  2  ep  0   adversary:  AdversaryModes.myopic
  actions:  tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22, 22, 22,
        22, 22, 22, 22, 22, 22,  0])
loss=  tensor(0.0355, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([10.5398, 10.5398, 10.5398,  9.8612,  9.2420,  8.6725,  8.1421,  7.6417,
         7.1642,  6.7037,  6.2560,  5.8179,  5.3870,  4.9614,  4.5399,  4.1213,
         3.7050,  3.2904,  2.8770,  2.4646,  2.0528,  1.6417,  1.2309,  0.8204,
         0.4101]) return=  168576.33348198733
probs of actions:  tensor([0.9992, 0.9984, 0.9995, 0.9995, 0.9995, 0.9996,
0.9996, 0.9998, 0.9994,
        0.9996, 0.9996, 0.9991, 0.9992, 0.9997, 0.9996, 0.9995, 0.9990, 0.9999,
```

```
        0.9998, 1.0000, 0.9996, 0.9999, 1.0000, 1.0000, 0.9956],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5718, 0.6302, 0.6604, 0.6757, 0.6834, 0.6873,
0.6893, 0.6902,
        0.6907, 0.6910, 0.6911, 0.6911, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912,
        0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.7396])
finalReturns:  tensor([5.2833, 5.3317, 5.2905, 5.1844, 5.0313, 4.8444, 4.6327,
4.4030, 4.1599,
        3.9071, 3.6469, 3.3813, 3.1117, 2.8391, 2.5642, 2.2876, 2.0098, 1.7310,
        1.4516, 1.1715, 0.8911, 0.6104, 0.3295])
----------------------------------------
iter  0  stage  1  ep  83   adversary:  AdversaryModes.myopic
  actions:  tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22, 22, 22,
        22, 22, 22, 22, 22, 22,  0])
loss=  tensor(0.0292, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([10.8367, 10.8367, 10.2165,  9.6255,  9.0696,  8.5458,  8.0486,  7.5726,
         7.1129,  6.6656,  6.2277,  5.7969,  5.3713,  4.9498,  4.5313,  4.1150,
         3.7003,  3.2870,  2.8745,  2.4628,  2.0516,  1.6408,  1.2303,  0.8201,
         0.4100]) return=  168576.33348198733
probs of actions:  tensor([0.9994, 0.9990, 0.9997, 0.9997, 0.9997, 0.9997,
0.9997, 0.9999, 0.9996,
        0.9998, 0.9997, 0.9994, 0.9995, 0.9998, 0.9998, 0.9997, 0.9992, 0.9999,
        0.9999, 1.0000, 0.9997, 1.0000, 1.0000, 1.0000, 0.9951],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5718, 0.6302, 0.6604, 0.6757, 0.6834, 0.6873,
0.6893, 0.6902,
        0.6907, 0.6910, 0.6911, 0.6911, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912,
        0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.7396])
finalReturns:  tensor([5.5581, 5.6065, 5.5674, 5.4630, 5.3111, 5.1248, 4.9135,
4.6840, 4.4410,
        4.1882, 3.9281, 3.6625, 3.3929, 3.1203, 2.8454, 2.5689, 2.2910, 2.0123,
        1.7328, 1.4528, 1.1724, 0.8917, 0.6107, 0.3296])
----------------------------------------
iter  0  stage  0  ep  0   adversary:  AdversaryModes.myopic
  actions:  tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22, 22, 22,
        22, 22, 22, 22, 22, 22,  0])
loss=  tensor(0.0333, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([11.0309, 10.5197,  9.9832,  9.4539,  8.9431,  8.4524,  7.9795,  7.5213,
         7.0747,  6.6372,  6.2066,  5.7811,  5.3597,  4.9411,  4.5248,  4.1102,
         3.6968,  3.2844,  2.8727,  2.4615,  2.0507,  1.6402,  1.2300,  0.8199,
         0.4099]) return=  168576.33348198733
probs of actions:  tensor([0.9994, 0.9990, 0.9997, 0.9997, 0.9997, 0.9997,
0.9997, 0.9999, 0.9996,
        0.9998, 0.9997, 0.9994, 0.9995, 0.9998, 0.9998, 0.9997, 0.9992, 0.9999,
        0.9999, 1.0000, 0.9997, 1.0000, 1.0000, 1.0000, 0.9951],
       grad_fn=<ExpBackward0>)
```

```
rewards:  tensor([0.4628, 0.5718, 0.6302, 0.6604, 0.6757, 0.6834, 0.6873,
0.6893, 0.6902,
        0.6907, 0.6910, 0.6911, 0.6911, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912,
        0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.6912, 0.7396])
finalReturns:  tensor([5.8267, 5.8751, 5.8399, 5.7390, 5.5894, 5.4045, 5.1940,
4.9648, 4.7221,
        4.4694, 4.2093, 3.9438, 3.6742, 3.4016, 3.1267, 2.8502, 2.5723, 2.2936,
        2.0141, 1.7341, 1.4537, 1.1730, 0.8920, 0.6109, 0.3297])
0,[1e-05,1][1, 10000, 1, 1],1682348121 saved
[346689, 'tensor([1., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.])',
168576.33348198733, 65680.3332469066, 0.03331249579787254, 1e-05, 1, 0,
'tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22,\n        22, 22, 22, 22, 22, 22,  0])', '[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.\n 1.]', '0,[1e-05,1][1, 10000, 1,
1],1682348121', 25, 50, 168576.3334819873, 205322.17620252006, 83998.798421993,
134475.68800000002, 131523.21866666665, 89197.96159612676, 89197.96159612676,
106340.34718300312, 106308.66365021843, 97483.02777699233, 89197.96159612676,
106308.66365021843]
policy reset
----------------------------------------
iter  1  stage  24  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0,
        0])
loss=  tensor(-0., grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.5625,
0.5625, 0.5625, 0.5625, 0.5625, 0.5625, 0.5625, 0.5625, 0.5625,
        0.5625, 0.5625, 0.5625, 0.5625, 0.5625, 0.5625, 0.5625, 0.5625, 0.5625,
        0.5625, 0.5625, 0.5625, 0.5625, 0.5625, 0.5625, 0.5625]) return=
139739.47699942059
probs of actions:  tensor([0.8929, 0.8857, 0.9020, 0.8962, 0.8926, 0.8984,
0.0646, 0.9006, 0.8831,
        0.9026, 0.8910, 0.0521, 0.8824, 0.8858, 0.8784, 0.9028, 0.8894, 0.8959,
        0.8758, 0.9148, 0.8829, 0.8869, 0.8875, 0.8934, 0.9822],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5112, 0.5366, 0.5495, 0.5560, 0.5592, 0.5609, 0.5616,
0.5658, 0.5642,
        0.5633, 0.5629, 0.5626, 0.5664, 0.5644, 0.5635, 0.5630, 0.5627, 0.5626,
        0.5626, 0.5625, 0.5625, 0.5625, 0.5625, 0.5625, 0.5625])
finalReturns:  tensor([0.])
----------------------------------------
iter  1  stage  23  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([ 1,  0, 11,  0,  0,  0, 16, 19,  0,  0,  0,  4, 14,  8,  0,
12, 11,  0,
        10,  4,  0,  0,  0, 14,  0])
loss=  tensor(0.0716, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.1330,
1.1330, 1.1330, 1.1330, 1.1330, 1.1330, 1.1330, 1.1330, 1.1330,
        1.1330, 1.1330, 1.1330, 1.1330, 1.1330, 1.1330, 1.1330, 1.1330, 1.1330,
        1.1330, 1.1330, 1.1330, 1.1330, 1.1330, 1.1330, 0.5652]) return=
```

```
146956.28276435166
probs of actions:  tensor([0.0785, 0.5092, 0.0653, 0.4961, 0.4847, 0.3295,
0.0636, 0.0042, 0.4223,
        0.5699, 0.5365, 0.0700, 0.0100, 0.0034, 0.3717, 0.0786, 0.0824, 0.3663,
        0.0485, 0.0709, 0.4142, 0.4306, 0.3190, 0.1280, 0.9781],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5111, 0.5402, 0.5392, 0.5987, 0.5805, 0.5714, 0.5414,
0.5903, 0.6695,
        0.6148, 0.5884, 0.5738, 0.5645, 0.6211, 0.6258, 0.5793, 0.6124, 0.6362,
        0.5888, 0.6176, 0.6060, 0.5840, 0.5732, 0.5482, 0.6190])
finalReturns:  tensor([0.0342, 0.0538])
----------------------------------------
iter  1  stage  22  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([16, 12, 28,  1,  0, 23, 12, 16, 26, 22,  0, 27, 16, 27, 12,
19, 19, 26,
         1, 27, 22, 30, 27, 20,  0])
loss=  tensor(0.9044, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.0149,
2.0149, 2.0149, 2.0149, 2.0149, 2.0149, 2.0149, 2.0149, 2.0149,
        2.0149, 2.0149, 2.0149, 2.0149, 2.0149, 2.0149, 2.0149, 2.0149, 2.0149,
        2.0149, 2.0149, 2.0149, 2.0149, 2.0149, 1.2464, 0.5849]) return=
162124.7477856951
probs of actions:  tensor([0.0491, 0.1345, 0.0154, 0.1258, 0.1692, 0.0112,
0.1468, 0.0460, 0.0490,
        0.0374, 0.2083, 0.0527, 0.0574, 0.0611, 0.2087, 0.0473, 0.0565, 0.0554,
        0.0369, 0.0536, 0.0429, 0.0126, 0.1198, 0.1008, 0.9626],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4856, 0.5824, 0.5477, 0.7066, 0.6365, 0.5460, 0.6571,
0.6382, 0.6087,
        0.6761, 0.7320, 0.5716, 0.6866, 0.6276, 0.7269, 0.6619, 0.6699, 0.6424,
        0.7418, 0.5802, 0.6683, 0.6381, 0.6956, 0.7358, 0.7489])
finalReturns:  tensor([0.1655, 0.2384, 0.1640])
----------------------------------------
iter  1  stage  21  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([29, 18, 27, 27, 26, 27, 29, 26, 27, 23, 27, 27, 19, 19, 20,
16, 25, 19,
        26, 23, 30, 27, 28, 29,  0])
loss=  tensor(3.3616, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.5627,
2.5627, 2.5627, 2.5627, 2.5627, 2.5627, 2.5627, 2.5627, 2.5627,
        2.5627, 2.5627, 2.5627, 2.5627, 2.5627, 2.5627, 2.5627, 2.5627, 2.5627,
        2.5627, 2.5627, 2.5627, 2.5627, 1.7836, 1.1172, 0.5300]) return=
170297.27457943128
probs of actions:  tensor([0.0463, 0.0392, 0.3955, 0.3738, 0.1785, 0.4352,
0.0531, 0.1506, 0.3757,
        0.0076, 0.4171, 0.4271, 0.0406, 0.0422, 0.0334, 0.0191, 0.0565, 0.0451,
        0.2052, 0.0074, 0.0165, 0.5405, 0.0105, 0.0368, 0.9833],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.6156, 0.6036, 0.6560, 0.6882, 0.6922, 0.6900,
0.7199, 0.7080,
```

```
        0.7292, 0.6922, 0.7012, 0.7426, 0.7099, 0.6899, 0.7006, 0.6449, 0.7001,
        0.6575, 0.6966, 0.6589, 0.7062, 0.7028, 0.7025, 0.7938])
finalReturns:  tensor([0.3426, 0.4155, 0.3792, 0.2638])
----------------------------------------
iter  1  stage  20  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([26, 27, 27, 25, 27, 27, 27,  0, 26, 27, 26, 27, 27, 27, 27,
27, 26, 25,
        26, 27, 27, 27, 27, 27,  0])
loss=  tensor(0.5544, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.0524,
3.0524, 3.0524, 3.0524, 3.0524, 3.0524, 3.0524, 3.0524, 3.0524,
        3.0524, 3.0524, 3.0524, 3.0524, 3.0524, 3.0524, 3.0524, 3.0524, 3.0524,
        3.0524, 3.0524, 3.0524, 2.2741, 1.6082, 1.0211, 0.4904]) return=
170770.7691723508
probs of actions:  tensor([0.1380, 0.6337, 0.6868, 0.0381, 0.6601, 0.7162,
0.7106, 0.0021, 0.1416,
        0.6234, 0.1196, 0.6613, 0.6336, 0.6924, 0.6360, 0.6668, 0.1431, 0.0319,
        0.1411, 0.7074, 0.7916, 0.8418, 0.8417, 0.6480, 0.9773],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4436, 0.5631, 0.6348, 0.6825, 0.6824, 0.6963, 0.7033,
0.7797, 0.5991,
        0.6466, 0.6834, 0.6898, 0.7000, 0.7052, 0.7077, 0.7090, 0.7150, 0.7160,
        0.7044, 0.7003, 0.7053, 0.7078, 0.7091, 0.7097, 0.7829])
finalReturns:  tensor([0.5625, 0.6354, 0.5935, 0.4715, 0.2926])
----------------------------------------
iter  1  stage  19  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([27, 27, 27, 29, 27, 27, 27, 27, 20, 27, 26, 27, 27, 27, 22,
27, 27, 27,
        27, 27, 27, 27, 27, 27,  0])
loss=  tensor(0.6967, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.5200,
3.5200, 3.5200, 3.5200, 3.5200, 3.5200, 3.5200, 3.5200, 3.5200,
        3.5200, 3.5200, 3.5200, 3.5200, 3.5200, 3.5200, 3.5200, 3.5200, 3.5200,
        3.5200, 3.5200, 2.7382, 2.0706, 1.4828, 0.9516, 0.4611]) return=
172225.21784845914
probs of actions:  tensor([0.7112, 0.6666, 0.7444, 0.0828, 0.7037, 0.7520,
0.7450, 0.6938, 0.0091,
        0.6783, 0.1017, 0.6977, 0.6331, 0.7429, 0.0189, 0.6977, 0.6746, 0.7301,
        0.7153, 0.8062, 0.8368, 0.8887, 0.8815, 0.7302, 0.9894],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5671, 0.6369, 0.6620, 0.7004, 0.7054, 0.7078,
0.7091, 0.7426,
        0.6794, 0.7001, 0.6981, 0.7042, 0.7073, 0.7333, 0.6876, 0.6989, 0.7046,
        0.7075, 0.7089, 0.7096, 0.7100, 0.7101, 0.7102, 0.7832])
finalReturns:  tensor([0.8120, 0.8849, 0.8429, 0.7208, 0.5418, 0.3221])
----------------------------------------
iter  1  stage  18  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([27, 27, 27, 29, 27, 26, 26, 27, 29, 29, 27, 27, 27, 27, 29,
27, 27, 27,
        27, 27, 27, 25, 27, 19,  0])
```

```
loss=  tensor(9.2171, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.9682,
3.9682, 3.9682, 3.9682, 3.9682, 3.9682, 3.9682, 3.9682, 3.9682,
        3.9682, 3.9682, 3.9682, 3.9682, 3.9682, 3.9682, 3.9682, 3.9682, 3.9682,
        3.9682, 3.1839, 2.5150, 1.9267, 1.3953, 0.9046, 0.4417]) return=
172834.79357956513
probs of actions:  tensor([0.7195, 0.6850, 0.7542, 0.0809, 0.7215, 0.1352,
0.1281, 0.7134, 0.0822,
        0.0737, 0.7545, 0.7188, 0.6647, 0.7534, 0.0863, 0.7126, 0.6873, 0.7424,
        0.7456, 0.8016, 0.8489, 0.0049, 0.8710, 0.0048, 0.9992],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5671, 0.6369, 0.6620, 0.7004, 0.7107, 0.7087,
0.7025, 0.6952,
        0.7060, 0.7227, 0.7165, 0.7134, 0.7119, 0.6999, 0.7196, 0.7149, 0.7126,
        0.7115, 0.7109, 0.7106, 0.7209, 0.7016, 0.7427, 0.7461])
finalReturns:  tensor([1.0760, 1.1489, 1.1069, 0.9846, 0.7951, 0.5842, 0.3044])
----------------------------------------
iter  1  stage  17  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([27, 27, 27, 27, 27, 27, 25, 27, 27, 27, 29, 27, 27, 27, 28,
26, 27, 27,
        26, 27, 27, 27, 27, 27,  0])
loss=  tensor(4.8350, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.3889,
4.3889, 4.3889, 4.3889, 4.3889, 4.3889, 4.3889, 4.3889, 4.3889,
        4.3889, 4.3889, 4.3889, 4.3889, 4.3889, 4.3889, 4.3889, 4.3889, 4.3889,
        3.6066, 2.9388, 2.3509, 1.8188, 1.3271, 0.8650, 0.4245]) return=
172634.51909284995
probs of actions:  tensor([0.7646, 0.7255, 0.7889, 0.7591, 0.7688, 0.8019,
0.0085, 0.7527, 0.7150,
        0.7432, 0.0773, 0.7542, 0.7013, 0.8001, 0.0441, 0.0799, 0.7366, 0.7465,
        0.0758, 0.8491, 0.8635, 0.9027, 0.9119, 0.8123, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5671, 0.6369, 0.6732, 0.6916, 0.7010, 0.7160,
0.6992, 0.7047,
        0.7075, 0.6977, 0.7185, 0.7144, 0.7124, 0.7058, 0.7206, 0.7084, 0.7093,
        0.7151, 0.7057, 0.7080, 0.7092, 0.7097, 0.7100, 0.7831])
finalReturns:  tensor([1.3613, 1.4342, 1.3869, 1.2691, 1.0932, 0.8757, 0.6281,
0.3586])
----------------------------------------
iter  1  stage  16  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([27, 29, 27, 27, 29, 27, 27, 27, 29, 27, 29, 28, 27, 27, 27,
27, 27, 27,
        27, 27, 27, 27, 27, 27,  0])
loss=  tensor(2.6017, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.7995,
4.7995, 4.7995, 4.7995, 4.7995, 4.7995, 4.7995, 4.7995, 4.7995,
        4.7995, 4.7995, 4.7995, 4.7995, 4.7995, 4.7995, 4.7995, 4.7995, 4.0157,
        3.3471, 2.7589, 2.2275, 1.7369, 1.2757, 0.8361, 0.4122]) return=
172966.5624334948
probs of actions:  tensor([0.7289, 0.1692, 0.7443, 0.7138, 0.1311, 0.7670,
0.7516, 0.7013, 0.1616,
```

```
        0.7016, 0.1237, 0.0473, 0.6433, 0.7727, 0.6639, 0.7063, 0.7170, 0.6757,
        0.8066, 0.8157, 0.8129, 0.8741, 0.9036, 0.8123, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5559, 0.6454, 0.6775, 0.6826, 0.7109, 0.7106,
0.7105, 0.6992,
        0.7192, 0.7036, 0.7159, 0.7203, 0.7153, 0.7128, 0.7116, 0.7109, 0.7106,
        0.7105, 0.7104, 0.7104, 0.7103, 0.7103, 0.7103, 0.7832])
finalReturns:  tensor([1.6675, 1.7404, 1.6984, 1.5761, 1.3971, 1.1774, 0.9282,
0.6575, 0.3710])
----------------------------------------
iter  1  stage  15  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([27, 29, 27, 27, 27, 29, 29, 27, 28, 29, 29, 27, 29, 27, 29,
27, 27, 29,
        29, 29, 27, 27, 27, 27,  0])
loss=  tensor(11.3290, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([5.1989, 5.1989, 5.1989, 5.1989, 5.1989, 5.1989, 5.1989, 5.1989, 5.1989,
        5.1989, 5.1989, 5.1989, 5.1989, 5.1989, 5.1989, 5.1989, 4.4036, 3.7298,
        3.1391, 2.6065, 2.1171, 1.6595, 1.2254, 0.8070, 0.3996]) return=
173351.18255934515
probs of actions:  tensor([0.5099, 0.3595, 0.5167, 0.4838, 0.5129, 0.3033,
0.3360, 0.4651, 0.0750,
        0.3347, 0.3011, 0.4653, 0.3972, 0.5718, 0.3602, 0.4803, 0.4625, 0.3816,
        0.2773, 0.3271, 0.6014, 0.6833, 0.7747, 0.6614, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5559, 0.6454, 0.6775, 0.6938, 0.6909, 0.7038,
0.7216, 0.7104,
        0.7064, 0.7116, 0.7255, 0.7067, 0.7230, 0.7054, 0.7224, 0.7163, 0.7021,
        0.7095, 0.7132, 0.7263, 0.7183, 0.7143, 0.7123, 0.7842])
finalReturns:  tensor([2.0200, 2.0929, 2.0505, 1.9390, 1.7621, 1.5383, 1.2696,
0.9855, 0.6895,
        0.3846])
----------------------------------------
iter  1  stage  14  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([29, 28, 28, 29, 29, 29, 29, 30, 27, 29, 26, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 26, 29,  0])
loss=  tensor(7.7630, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.5612,
5.5612, 5.5612, 5.5612, 5.5612, 5.5612, 5.5612, 5.5612, 5.5612, 5.5612,
        5.5612, 5.5612, 5.5612, 5.5612, 5.5612, 5.5612, 4.7621, 4.0865, 3.4969,
        2.9681, 2.4827, 2.0287, 1.5976, 1.1832, 0.7812, 0.3884]) return=
173825.5551842266
probs of actions:  tensor([0.7370, 0.0655, 0.0614, 0.7672, 0.7414, 0.7641,
0.7839, 0.0485, 0.0853,
        0.7527, 0.0260, 0.7726, 0.7826, 0.7109, 0.7644, 0.8093, 0.8081, 0.7662,
        0.7728, 0.8169, 0.8252, 0.7849, 0.0323, 0.6756, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5696, 0.6399, 0.6706, 0.6936, 0.7052, 0.7111,
0.7081, 0.7311,
```

```
        0.7095, 0.7297, 0.7017, 0.7093, 0.7131, 0.7150, 0.7160, 0.7164, 0.7167,
        0.7168, 0.7169, 0.7169, 0.7169, 0.7334, 0.7036, 0.7943])
finalReturns:  tensor([2.4016, 2.4857, 2.4454, 2.3186, 2.1307, 1.8993, 1.6364,
1.3506, 1.0481,
        0.7166, 0.4059])
---------------------------------------
iter  1  stage  13  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([29, 29, 29, 29, 29, 29, 30, 29, 28, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 27, 29, 29, 29,  0])
loss=  tensor(8.1818, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.9484,
5.9484, 5.9484, 5.9484, 5.9484, 5.9484, 5.9484, 5.9484, 5.9484,
        5.9484, 5.9484, 5.9484, 5.9484, 5.9484, 5.1476, 4.4712, 3.8812, 3.3522,
        2.8668, 2.4127, 1.9816, 1.5672, 1.1652, 0.7708, 0.3828]) return=
174002.82550723848
probs of actions:  tensor([0.7929, 0.8025, 0.8118, 0.8192, 0.7962, 0.8142,
0.0197, 0.8157, 0.0652,
        0.8068, 0.8006, 0.8221, 0.8335, 0.7760, 0.8249, 0.8528, 0.8706, 0.7983,
        0.8249, 0.8632, 0.0896, 0.8470, 0.7300, 0.7179, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5639, 0.6384, 0.6772, 0.6969, 0.7069, 0.7060,
0.7189, 0.7236,
        0.7129, 0.7149, 0.7159, 0.7164, 0.7167, 0.7168, 0.7169, 0.7169, 0.7169,
        0.7169, 0.7169, 0.7281, 0.7080, 0.7125, 0.7147, 0.7999])
finalReturns:  tensor([2.7328, 2.8169, 2.7765, 2.6496, 2.4617, 2.2302, 1.9674,
1.6816, 1.3678,
        1.0618, 0.7438, 0.4171])
---------------------------------------
iter  1  stage  12  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([29, 29, 29, 29, 28, 29, 29, 29, 29, 29, 29, 30, 29, 27, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(12.7920, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([6.3411, 6.3411, 6.3411, 6.3411, 6.3411, 6.3411, 6.3411, 6.3411, 6.3411,
        6.3411, 6.3411, 6.3411, 6.3411, 5.5357, 4.8572, 4.2662, 3.7349, 3.2471,
        2.7909, 2.3581, 1.9423, 1.5393, 1.1457, 0.7590, 0.3776]) return=
173992.11630517073
probs of actions:  tensor([0.8624, 0.8648, 0.8793, 0.8813, 0.0524, 0.8886,
0.8990, 0.8753, 0.8682,
        0.8730, 0.8802, 0.0325, 0.8614, 0.0400, 0.8668, 0.9117, 0.9162, 0.8509,
        0.9006, 0.9281, 0.9482, 0.9433, 0.8928, 0.8473, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5639, 0.6384, 0.6772, 0.7026, 0.7024, 0.7097,
0.7133, 0.7151,
        0.7160, 0.7165, 0.7108, 0.7213, 0.7303, 0.7091, 0.7130, 0.7150, 0.7159,
        0.7164, 0.7167, 0.7168, 0.7169, 0.7169, 0.7169, 0.8010])
finalReturns:  tensor([3.0651, 3.1492, 3.0974, 2.9793, 2.7976, 2.5705, 2.3107,
2.0271, 1.7261,
```

```
        1.4124, 1.0892, 0.7589, 0.4235])
--------------------------------------
iter  1  stage  11  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([29, 29, 29, 29, 30, 27, 29, 30, 30, 28, 29, 28, 29, 29, 29,
29, 29, 30,
        30, 29, 29, 29, 29, 29,  0])
loss=  tensor(27.1738, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([6.6912, 6.6912, 6.6912, 6.6912, 6.6912, 6.6912, 6.6912, 6.6912, 6.6912,
        6.6912, 6.6912, 6.6912, 5.8911, 5.2150, 4.6242, 4.0942, 3.6077, 3.1528,
        2.7210, 2.3062, 1.9046, 1.5132, 1.1288, 0.7493, 0.3734]) return=
174079.33224464272
probs of actions:  tensor([0.8209, 0.8254, 0.8419, 0.8423, 0.0719, 0.0270,
0.8679, 0.0696, 0.0810,
        0.0624, 0.8521, 0.0596, 0.7923, 0.8248, 0.8013, 0.8720, 0.8975, 0.0851,
        0.0489, 0.8874, 0.9414, 0.9314, 0.8985, 0.8402, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5639, 0.6384, 0.6772, 0.6910, 0.7225, 0.7052,
0.7052, 0.7126,
        0.7279, 0.7151, 0.7217, 0.7120, 0.7145, 0.7157, 0.7163, 0.7166, 0.7109,
        0.7154, 0.7236, 0.7203, 0.7186, 0.7178, 0.7173, 0.8012])
finalReturns:  tensor([3.4307, 3.5091, 3.4732, 3.3495, 3.1639, 2.9340, 2.6723,
2.3932, 2.0926,
        1.7706, 1.4418, 1.1075, 0.7693, 0.4279])
--------------------------------------
iter  1  stage  10  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([29, 29, 26, 29, 29, 29, 29, 29, 29, 28, 29, 29, 28, 29, 30,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(21.4632, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([7.0587, 7.0587, 7.0587, 7.0587, 7.0587, 7.0587, 7.0587, 7.0587, 7.0587,
        7.0587, 7.0587, 6.2626, 5.5883, 4.9994, 4.4709, 3.9847, 3.5298, 3.0985,
        2.6844, 2.2827, 1.8901, 1.5043, 1.1236, 0.7465, 0.3722]) return=
173927.67848376333
probs of actions:  tensor([0.7494, 0.7613, 0.0041, 0.7772, 0.7584, 0.8050,
0.8112, 0.7614, 0.7421,
        0.0720, 0.8168, 0.7777, 0.0912, 0.7630, 0.1784, 0.7979, 0.8219, 0.7332,
        0.8220, 0.8517, 0.9236, 0.9042, 0.8870, 0.8024, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5639, 0.6549, 0.6641, 0.6903, 0.7036, 0.7102,
0.7136, 0.7152,
        0.7218, 0.7120, 0.7145, 0.7214, 0.7118, 0.7085, 0.7201, 0.7185, 0.7177,
        0.7173, 0.7171, 0.7170, 0.7170, 0.7170, 0.7169, 0.8010])
finalReturns:  tensor([3.7693, 3.8534, 3.8131, 3.6807, 3.4974, 3.2750, 3.0098,
2.7226, 2.4190,
        2.1033, 1.7788, 1.4476, 1.1114, 0.7715, 0.4288])
--------------------------------------
iter  1  stage  9  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([29, 29, 29, 30, 30, 29, 30, 29, 29, 29, 29, 29, 29, 29, 29,
```

```
29, 29, 28,
        29, 29, 29, 29, 29, 28,  0])
loss=  tensor(16.1739, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([7.4444, 7.4444, 7.4444, 7.4444, 7.4444, 7.4444, 7.4444, 7.4444, 7.4444,
        7.4444, 6.6425, 5.9655, 5.3753, 4.8462, 4.3607, 3.9066, 3.4755, 3.0611,
        2.6591, 2.2663, 1.8794, 1.4975, 1.1193, 0.7442, 0.3712]) return=
174131.5237391631
probs of actions:  tensor([0.8230, 0.8277, 0.8443, 0.0909, 0.0966, 0.8673,
0.0717, 0.8321, 0.8166,
        0.8434, 0.8727, 0.8568, 0.7877, 0.8095, 0.7937, 0.8621, 0.8774, 0.0635,
        0.8763, 0.9007, 0.9587, 0.9444, 0.9358, 0.0460, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5639, 0.6384, 0.6713, 0.6954, 0.7136, 0.7093,
0.7206, 0.7187,
        0.7178, 0.7174, 0.7172, 0.7170, 0.7170, 0.7170, 0.7169, 0.7169, 0.7226,
        0.7125, 0.7147, 0.7158, 0.7164, 0.7166, 0.7225, 0.7965])
finalReturns:  tensor([4.1104, 4.1945, 4.1540, 4.0271, 3.8392, 3.6077, 3.3448,
3.0590, 2.7564,
        2.4358, 2.1162, 1.7883, 1.4545, 1.1163, 0.7748, 0.4253])
----------------------------------------
iter  1  stage  8  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 28, 29, 29, 29, 30, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(30.2489, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([7.8061, 7.8061, 7.8061, 7.8061, 7.8061, 7.8061, 7.8061, 7.8061, 7.8061,
        7.0063, 6.3304, 5.7406, 5.2108, 4.7243, 4.2693, 3.8374, 3.4231, 3.0216,
        2.6293, 2.2438, 1.8632, 1.4863, 1.1121, 0.7400, 0.3694]) return=
174058.37645850837
probs of actions:  tensor([0.8582, 0.8629, 0.8798, 0.8764, 0.8638, 0.9002,
0.9004, 0.8670, 0.8585,
        0.0301, 0.9124, 0.8797, 0.8310, 0.0907, 0.8367, 0.8905, 0.9087, 0.8720,
        0.9059, 0.9327, 0.9686, 0.9607, 0.9600, 0.9044, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5639, 0.6384, 0.6772, 0.6969, 0.7069, 0.7119,
0.7144, 0.7157,
        0.7220, 0.7121, 0.7145, 0.7157, 0.7104, 0.7211, 0.7190, 0.7180, 0.7174,
        0.7172, 0.7171, 0.7170, 0.7170, 0.7169, 0.7169, 0.8010])
finalReturns:  tensor([4.4630, 4.5471, 4.5011, 4.3787, 4.1940, 3.9647, 3.7093,
3.4202, 3.1154,
        2.7989, 2.4738, 2.1422, 1.8057, 1.4656, 1.1228, 0.7779, 0.4316])
----------------------------------------
iter  1  stage  7  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(4.4280, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.1717,
8.1717, 8.1717, 8.1717, 8.1717, 8.1717, 8.1717, 8.1717, 7.3731,
```

```
        6.6978, 6.1083, 5.5795, 5.0942, 4.6402, 4.2091, 3.7948, 3.3928, 2.9999,
        2.6139, 2.2329, 1.8557, 1.4813, 1.1089, 0.7382, 0.3687]) return=
174060.25019204617
probs of actions:  tensor([0.9034, 0.9041, 0.9183, 0.9161, 0.9079, 0.9342,
0.9338, 0.9166, 0.9100,
        0.9035, 0.9549, 0.9257, 0.8960, 0.9082, 0.8750, 0.9167, 0.9458, 0.9095,
        0.9417, 0.9605, 0.9816, 0.9780, 0.9770, 0.9405, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5639, 0.6384, 0.6772, 0.6969, 0.7069, 0.7119,
0.7144, 0.7157,
        0.7163, 0.7166, 0.7168, 0.7168, 0.7169, 0.7169, 0.7169, 0.7169, 0.7169,
        0.7169, 0.7169, 0.7169, 0.7169, 0.7169, 0.7169, 0.7169, 0.8010])
finalReturns:  tensor([4.8121, 4.8962, 4.8558, 4.7290, 4.5412, 4.3098, 4.0469,
3.7611, 3.4586,
        3.1436, 2.8196, 2.4887, 2.1528, 1.8131, 1.4705, 1.1259, 0.7797, 0.4324])
----------------------------------------
iter  1  stage  6  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        28, 29, 29, 29, 29, 29,  0])
loss=  tensor(17.8648, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([8.5390, 8.5390, 8.5390, 8.5390, 8.5390, 8.5390, 8.5390, 7.7430, 7.0688,
        6.4798, 5.9513, 5.4661, 5.0122, 4.5811, 4.1668, 3.7648, 3.3719, 2.9859,
        2.6049, 2.2277, 1.8533, 1.4802, 1.1085, 0.7381, 0.3686]) return=
174029.23196541704
probs of actions:  tensor([0.8938, 0.8960, 0.9100, 0.9065, 0.9003, 0.9289,
0.9158, 0.9051, 0.9186,
        0.8903, 0.9594, 0.9238, 0.8900, 0.9022, 0.8511, 0.9066, 0.9395, 0.9038,
        0.0062, 0.9561, 0.9805, 0.9749, 0.9774, 0.9416, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5639, 0.6384, 0.6772, 0.6969, 0.7069, 0.7119,
0.7144, 0.7157,
        0.7163, 0.7166, 0.7168, 0.7168, 0.7169, 0.7169, 0.7169, 0.7169, 0.7169,
        0.7226, 0.7125, 0.7147, 0.7158, 0.7164, 0.7166, 0.8009])
finalReturns:  tensor([5.1535, 5.2376, 5.1974, 5.0707, 4.8829, 4.6515, 4.3887,
4.1029, 3.8004,
        3.4854, 3.1614, 2.8305, 2.4946, 2.1492, 1.8111, 1.4695, 1.1254, 0.7795,
        0.4323])
----------------------------------------
iter  1  stage  5  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(2.6399, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.8931,
8.8931, 8.8931, 8.8931, 8.8931, 8.8931, 8.1021, 7.4302, 6.8424,
        6.3143, 5.8294, 5.3756, 4.9445, 4.5302, 4.1283, 3.7354, 3.3494, 2.9684,
        2.5912, 2.2168, 1.8445, 1.4737, 1.1042, 0.7355, 0.3675]) return=
174060.25019204617
```

```
probs of actions:  tensor([0.9462, 0.9455, 0.9555, 0.9534, 0.9488, 0.9645,
0.9683, 0.9549, 0.9581,
        0.9596, 0.9844, 0.9700, 0.9389, 0.9530, 0.9245, 0.9614, 0.9708, 0.9519,
        0.9705, 0.9807, 0.9920, 0.9906, 0.9916, 0.9737, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5639, 0.6384, 0.6772, 0.6969, 0.7069, 0.7119,
0.7144, 0.7157,
        0.7163, 0.7166, 0.7168, 0.7168, 0.7169, 0.7169, 0.7169, 0.7169, 0.7169,
        0.7169, 0.7169, 0.7169, 0.7169, 0.7169, 0.7169, 0.8010])
finalReturns:  tensor([5.5094, 5.5935, 5.5535, 5.4270, 5.2393, 5.0080, 4.7452,
4.4594, 4.1569,
        3.8420, 3.5179, 3.1870, 2.8511, 2.5114, 2.1689, 1.8243, 1.4781, 1.1307,
        0.7824, 0.4335])
----------------------------------------
iter  1  stage  4  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 30, 29, 29, 29, 29, 29, 30, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(32.6679, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([9.2297, 9.2297, 9.2297, 9.2297, 9.2297, 8.4486, 7.7814, 7.1957, 6.6687,
        6.1842, 5.7315, 5.3016, 4.8883, 4.4871, 4.0950, 3.7094, 3.3296, 2.9536,
        2.5804, 2.2090, 1.8390, 1.4701, 1.1018, 0.7342, 0.3669]) return=
174121.3274714346
probs of actions:  tensor([0.9370, 0.9376, 0.9482, 0.9457, 0.9450, 0.9597,
0.9723, 0.0570, 0.9488,
        0.9538, 0.9850, 0.9677, 0.9243, 0.0438, 0.9097, 0.9542, 0.9587, 0.9397,
        0.9704, 0.9754, 0.9915, 0.9905, 0.9908, 0.9710, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5639, 0.6384, 0.6772, 0.6969, 0.7069, 0.7119,
0.7085, 0.7201,
        0.7185, 0.7177, 0.7173, 0.7171, 0.7111, 0.7215, 0.7192, 0.7181, 0.7175,
        0.7172, 0.7171, 0.7170, 0.7170, 0.7169, 0.7169, 0.8010])
finalReturns:  tensor([5.8759, 5.9600, 5.9204, 5.7942, 5.6126, 5.3770, 5.1112,
4.8233, 4.5193,
        4.2033, 3.8844, 3.5484, 3.2091, 2.8670, 2.5228, 2.1769, 1.8298, 1.4818,
        1.1331, 0.7838, 0.4341])
----------------------------------------
iter  1  stage  3  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([29, 29, 30, 29, 29, 29, 29, 30, 29, 29, 29, 29, 30, 29, 29,
29, 30, 29,
        30, 29, 29, 29, 29, 29,  0])
loss=  tensor(41.6779, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([9.5587, 9.5587, 9.5587, 9.5587, 8.7930, 8.1329, 7.5505, 7.0251, 6.5414,
        6.0882, 5.6583, 5.2451, 4.8442, 4.4522, 4.0668, 3.6871, 3.3112, 2.9380,
        2.5667, 2.1975, 1.8295, 1.4630, 1.0971, 0.7314, 0.3657]) return=
174209.913456734
probs of actions:  tensor([0.8751, 0.8809, 0.1046, 0.8975, 0.8943, 0.9066,
0.9349, 0.1239, 0.8953,
```

```
        0.9055, 0.9732, 0.9129, 0.1605, 0.9055, 0.8477, 0.9023, 0.0858, 0.8891,
        0.0721, 0.9442, 0.9840, 0.9776, 0.9772, 0.9326, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5639, 0.6325, 0.6815, 0.6991, 0.7080, 0.7125,
0.7088, 0.7203,
        0.7186, 0.7178, 0.7173, 0.7112, 0.7215, 0.7192, 0.7181, 0.7116, 0.7217,
        0.7134, 0.7226, 0.7198, 0.7183, 0.7176, 0.7173, 0.8012])
finalReturns:  tensor([6.2388, 6.3229, 6.2839, 6.1583, 5.9712, 5.7461, 5.4791,
5.1904, 4.8858,
        4.5694, 4.2501, 3.9140, 3.5745, 3.2323, 2.8939, 2.5435, 2.1993, 1.8447,
        1.4914, 1.1390, 0.7871, 0.4355])
----------------------------------------
iter  1  stage  2  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([29, 29, 29, 29, 29, 29, 30, 29, 29, 30, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 30,  0])
loss=  tensor(40.5944, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([9.8524, 9.8524, 9.8524, 9.1299, 8.4899, 7.9170, 7.3961, 6.9146, 6.4624,
        6.0330, 5.6201, 5.2193, 4.8281, 4.4437, 4.0642, 3.6881, 3.3147, 2.9431,
        2.5729, 2.2038, 1.8354, 1.4677, 1.1003, 0.7333, 0.3666]) return=
174107.14165860406
probs of actions:  tensor([0.8416, 0.8520, 0.8595, 0.8623, 0.8794, 0.8895,
0.0793, 0.8476, 0.8646,
        0.1119, 0.9667, 0.8832, 0.7889, 0.8842, 0.7762, 0.8842, 0.8902, 0.8377,
        0.9165, 0.9302, 0.9792, 0.9674, 0.9688, 0.0917, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5639, 0.6384, 0.6772, 0.6969, 0.7069, 0.7060,
0.7189, 0.7179,
        0.7115, 0.7217, 0.7193, 0.7181, 0.7175, 0.7172, 0.7171, 0.7170, 0.7170,
        0.7169, 0.7169, 0.7169, 0.7169, 0.7169, 0.7110, 0.8055])
finalReturns:  tensor([6.5673, 6.6514, 6.6142, 6.4902, 6.3042, 6.0797, 5.8130,
5.5245, 5.2259,
        4.9051, 4.5769, 4.2432, 3.9053, 3.5641, 3.2205, 2.8751, 2.5283, 2.1805,
        1.8319, 1.4827, 1.1331, 0.7832, 0.4389])
----------------------------------------
iter  1  stage  1  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([30, 30, 30, 29, 29, 29, 29, 30, 29, 29, 29, 29, 29, 30, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(54.3465, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([10.0714, 10.0714,  9.4193,  8.8129,  8.2569,  7.7457,  7.2699,  6.8214,
         6.3935,  5.9812,  5.5814,  5.1904,  4.8059,  4.4262,  4.0499,  3.6762,
         3.3052,  2.9359,  2.5675,  2.1999,  1.8327,  1.4658,  1.0991,  0.7327,
         0.3663]) return=  174202.79502442433
probs of actions:  tensor([0.2038, 0.1953, 0.1608, 0.8084, 0.8417, 0.8700,
0.8805, 0.2281, 0.8222,
        0.8522, 0.9632, 0.8519, 0.6879, 0.1521, 0.7490, 0.8204, 0.8533, 0.7903,
        0.8795, 0.8923, 0.9703, 0.9507, 0.9572, 0.8827, 1.0000],
```

```
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4212, 0.5621, 0.6389, 0.6848, 0.7008, 0.7088, 0.7129,
0.7090, 0.7204,
        0.7187, 0.7178, 0.7174, 0.7171, 0.7111, 0.7215, 0.7192, 0.7181, 0.7175,
        0.7172, 0.7171, 0.7170, 0.7170, 0.7169, 0.7169, 0.8010])
finalReturns:  tensor([6.9277, 7.0177, 6.9853, 6.8564, 6.6668, 6.4338, 6.1694,
5.8883, 5.5802,
        5.2614, 4.9345, 4.6017, 4.2643, 3.9294, 3.5817, 3.2335, 2.8847, 2.5356,
        2.1860, 1.8362, 1.4861, 1.1358, 0.7853, 0.4347])
----------------------------------------
iter  1  stage  0  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([30, 29, 29, 30, 30, 29, 29, 29, 29, 30, 29, 29, 30, 29, 29,
29, 29, 29,
        29, 30, 29, 30, 29, 29,  0])
loss=  tensor(69.6359, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([10.1501,  9.6388,  9.1023,  8.5803,  8.0846,  7.6157,  7.1709,  6.7459,
         6.3358,  5.9369,  5.5462,  5.1617,  4.7827,  4.4073,  4.0344,  3.6641,
         3.2953,  2.9274,  2.5601,  2.1932,  1.8265,  1.4600,  1.0944,  0.7290,
         0.3645]) return=  174253.3722751547
probs of actions:  tensor([0.2702, 0.7677, 0.7977, 0.2116, 0.2319, 0.8234,
0.8262, 0.7327, 0.7965,
        0.1930, 0.9485, 0.7959, 0.3707, 0.8006, 0.6535, 0.7538, 0.8028, 0.7528,
        0.8439, 0.1603, 0.9571, 0.0712, 0.9422, 0.8500, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4212, 0.5680, 0.6405, 0.6723, 0.6960, 0.7139, 0.7154,
0.7162, 0.7165,
        0.7108, 0.7213, 0.7191, 0.7121, 0.7220, 0.7194, 0.7182, 0.7176, 0.7172,
        0.7171, 0.7111, 0.7214, 0.7133, 0.7225, 0.7197, 0.8024])
finalReturns:  tensor([7.2753, 7.3653, 7.3339, 7.2154, 7.0387, 6.8116, 6.5426,
6.2521, 5.9461,
        5.6285, 5.3083, 4.9715, 4.6314, 4.2947, 3.9456, 3.5965, 3.2471, 2.8974,
        2.5475, 2.1973, 1.8529, 1.4980, 1.1503, 0.7931, 0.4380])
0,[1e-05,1][1, 10000, 1, 1],1682423487 saved
[3000000, 'tensor([1., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.])',
174210.39338284553, 57605.93776875555, 65.17557525634766, 1e-05, 1, 0,
'tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 30, 30, 30, 29, 29, 29, 30, 29,
29,\n        29, 30, 29, 29, 29, 29,  0])', '[0.69 0.75 0.77 0.78 0.75 0.8  0.8
0.7  0.77 0.2  0.06 0.24 0.6  0.77\n 0.62 0.27 0.78 0.72 0.82 0.18 0.95 0.92
0.94 0.84 1.  ]', '0,[1e-05,1][1, 10000, 1, 1],1682423487', 25, 50,
174199.95044967832, 226157.05867704182, 94851.05074168817, 131012.56797668608,
127973.03513660273, 64641.60648389723, 62848.849838023714, 78979.431849868,
79951.69168142487, 109515.23673882235, 64159.865982403535, 79875.30170982817]
policy reset
----------------------------------------
iter  2  stage  24  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([0, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 6, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0,
        0])
```

```
loss=  tensor(-0., grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.5625,
0.5625, 0.5625, 0.5625, 0.5625, 0.5625, 0.5625, 0.5625, 0.5625,
        0.5625, 0.5625, 0.5625, 0.5625, 0.5625, 0.5625, 0.5625, 0.5625, 0.5625,
        0.5625, 0.5625, 0.5625, 0.5625, 0.5625, 0.5625, 0.5625]) return=
140290.89393391204
probs of actions:  tensor([0.9195, 0.0062, 0.9061, 0.9295, 0.9160, 0.9197,
0.9265, 0.9012, 0.8971,
        0.9135, 0.9181, 0.9058, 0.9080, 0.9095, 0.0020, 0.9082, 0.9207, 0.9140,
        0.9216, 0.8805, 0.9001, 0.9244, 0.8986, 0.9275, 0.9888],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5112, 0.5350, 0.5644, 0.5634, 0.5630, 0.5627, 0.5626,
0.5626, 0.5625,
        0.5625, 0.5625, 0.5625, 0.5625, 0.5625, 0.5589, 0.5852, 0.5738, 0.5681,
        0.5653, 0.5639, 0.5632, 0.5629, 0.5627, 0.5626, 0.5625])
finalReturns:  tensor([0.])
----------------------------------------
iter  2  stage  23  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([17,  0,  1, 22,  8,  0,  9, 17, 19,  7, 14,  7,  0,  1, 14,
0,  7,  1,
         0, 10, 17,  0,  2, 21,  0])
loss=  tensor(0.1059, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.1688,
1.1688, 1.1688, 1.1688, 1.1688, 1.1688, 1.1688, 1.1688, 1.1688,
        1.1688, 1.1688, 1.1688, 1.1688, 1.1688, 1.1688, 1.1688, 1.1688, 1.1688,
        1.1688, 1.1688, 1.1688, 1.1688, 1.1688, 1.1688, 0.5770]) return=
150673.14503249113
probs of actions:  tensor([0.0195, 0.3042, 0.0644, 0.0075, 0.0056, 0.3819,
0.0345, 0.0280, 0.0135,
        0.0520, 0.0902, 0.0551, 0.2481, 0.0579, 0.0623, 0.3358, 0.0548, 0.0627,
        0.5367, 0.0444, 0.0219, 0.4604, 0.0402, 0.0654, 0.9866],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4823, 0.6006, 0.5813, 0.5273, 0.6487, 0.6395, 0.5923,
0.5872, 0.6199,
        0.6798, 0.6304, 0.6563, 0.6385, 0.5998, 0.5653, 0.6279, 0.5898, 0.6053,
        0.5876, 0.5650, 0.5782, 0.6514, 0.6057, 0.5477, 0.6596])
finalReturns:  tensor([0.0384, 0.0825])
----------------------------------------
iter  2  stage  22  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([23, 19, 13, 24, 23,  1, 23, 25, 24, 24, 23, 17, 24, 24, 13,
7, 21, 16,
        23, 23, 24, 23, 24, 24,  0])
loss=  tensor(0.3284, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.9812,
1.9812, 1.9812, 1.9812, 1.9812, 1.9812, 1.9812, 1.9812, 1.9812,
        1.9812, 1.9812, 1.9812, 1.9812, 1.9812, 1.9812, 1.9812, 1.9812, 1.9812,
        1.9812, 1.9812, 1.9812, 1.9812, 1.9812, 1.2344, 0.5830]) return=
165779.32260088876
probs of actions:  tensor([0.1627, 0.0385, 0.0941, 0.1479, 0.1835, 0.0145,
0.1909, 0.0031, 0.2096,
        0.1863, 0.2033, 0.2533, 0.2315, 0.1846, 0.2161, 0.0282, 0.0494, 0.0190,
```

```
        0.1273, 0.1832, 0.1901, 0.1711, 0.5221, 0.3500, 0.9898],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4583, 0.5880, 0.6514, 0.6087, 0.6580, 0.7293, 0.5944,
0.6343, 0.6732,
        0.6862, 0.6974, 0.7204, 0.6654, 0.6823, 0.7315, 0.7008, 0.6161, 0.6696,
        0.6391, 0.6670, 0.6764, 0.6925, 0.6892, 0.6942, 0.7544])
finalReturns:  tensor([0.1566, 0.2142, 0.1714])
----------------------------------------
iter  2  stage  21  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([24, 24, 24, 19, 24, 24, 24, 24, 17, 24, 24, 24, 24, 24, 24,
24, 24, 24,
        24, 24, 24, 24, 24, 23,  0])
loss=  tensor(0.6763, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.5329,
2.5329, 2.5329, 2.5329, 2.5329, 2.5329, 2.5329, 2.5329, 2.5329,
        2.5329, 2.5329, 2.5329, 2.5329, 2.5329, 2.5329, 2.5329, 2.5329, 2.5329,
        2.5329, 2.5329, 2.5329, 2.5329, 1.7760, 1.1199, 0.5347]) return=
169798.62492291527
probs of actions:  tensor([0.7248, 0.6904, 0.7508, 0.0141, 0.7384, 0.7358,
0.8043, 0.6820, 0.0194,
        0.7494, 0.7018, 0.6975, 0.7651, 0.7216, 0.7007, 0.7041, 0.7644, 0.6917,
        0.7048, 0.7520, 0.7286, 0.8208, 0.8729, 0.1889, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4536, 0.5705, 0.6334, 0.6875, 0.6612, 0.6801, 0.6897,
0.6945, 0.7256,
        0.6680, 0.6836, 0.6914, 0.6953, 0.6973, 0.6983, 0.6988, 0.6991, 0.6992,
        0.6992, 0.6993, 0.6993, 0.6993, 0.6993, 0.7040, 0.7526])
finalReturns:  tensor([0.3222, 0.3798, 0.3366, 0.2178])
----------------------------------------
iter  2  stage  20  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([24, 24, 28, 24, 23, 16, 24, 28, 24, 24, 24, 24, 24, 28, 23,
23, 24, 23,
        24, 23, 24, 24, 24, 24,  0])
loss=  tensor(0.6225, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.0211,
3.0211, 3.0211, 3.0211, 3.0211, 3.0211, 3.0211, 3.0211, 3.0211,
        3.0211, 3.0211, 3.0211, 3.0211, 3.0211, 3.0211, 3.0211, 3.0211, 3.0211,
        3.0211, 3.0211, 3.0211, 2.2697, 1.6162, 1.0322, 0.4981]) return=
170179.017451197
probs of actions:  tensor([0.6627, 0.6456, 0.1464, 0.6529, 0.1292, 0.0016,
0.7567, 0.1929, 0.6563,
        0.6988, 0.6466, 0.6345, 0.7108, 0.1598, 0.1356, 0.1445, 0.7143, 0.1483,
        0.6384, 0.1183, 0.6669, 0.7142, 0.8405, 0.7487, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4536, 0.5705, 0.6126, 0.6831, 0.6959, 0.7229, 0.6608,
0.6591, 0.7070,
        0.7031, 0.7012, 0.7003, 0.6998, 0.6787, 0.7216, 0.7084, 0.6972, 0.7029,
        0.6944, 0.7016, 0.6937, 0.6965, 0.6979, 0.6986, 0.7566])
finalReturns:  tensor([0.5223, 0.5799, 0.5369, 0.4229, 0.2585])
----------------------------------------
```

```
iter  2  stage  19  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28,
28, 28, 24,
        28, 28, 28, 28, 28, 24,  0])
loss=  tensor(1.3065, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.5129,
3.5129, 3.5129, 3.5129, 3.5129, 3.5129, 3.5129, 3.5129, 3.5129,
        3.5129, 3.5129, 3.5129, 3.5129, 3.5129, 3.5129, 3.5129, 3.5129, 3.5129,
        3.5129, 3.5129, 2.7297, 2.0614, 1.4742, 0.9450, 0.4574]) return=
173274.47945104944
probs of actions:  tensor([0.8751, 0.8362, 0.8357, 0.8808, 0.8721, 0.8655,
0.8324, 0.8766, 0.8592,
        0.8549, 0.8751, 0.8901, 0.8497, 0.8566, 0.8846, 0.8587, 0.8459, 0.0844,
        0.8948, 0.8662, 0.9174, 0.9180, 0.8195, 0.2077, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4328, 0.5656, 0.6377, 0.6752, 0.6944, 0.7040, 0.7088,
0.7113, 0.7125,
        0.7131, 0.7134, 0.7135, 0.7136, 0.7137, 0.7137, 0.7137, 0.7137, 0.7345,
        0.6960, 0.7048, 0.7093, 0.7115, 0.7126, 0.7339, 0.7741])
finalReturns:  tensor([0.8333, 0.9117, 0.8708, 0.7464, 0.5631, 0.3168])
----------------------------------------
iter  2  stage  18  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28,
28, 28, 28,
        28, 28, 28, 28, 28, 28,  0])
loss=  tensor(0.1852, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.9647,
3.9647, 3.9647, 3.9647, 3.9647, 3.9647, 3.9647, 3.9647, 3.9647,
        3.9647, 3.9647, 3.9647, 3.9647, 3.9647, 3.9647, 3.9647, 3.9647, 3.9647,
        3.9647, 3.1726, 2.5002, 1.9112, 1.3810, 0.8930, 0.4354]) return=
173387.3335190018
probs of actions:  tensor([0.9707, 0.9563, 0.9583, 0.9732, 0.9703, 0.9683,
0.9593, 0.9722, 0.9662,
        0.9654, 0.9716, 0.9762, 0.9634, 0.9665, 0.9740, 0.9665, 0.9630, 0.9683,
        0.9812, 0.9737, 0.9835, 0.9850, 0.9637, 0.8965, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4328, 0.5656, 0.6377, 0.6752, 0.6944, 0.7040, 0.7088,
0.7113, 0.7125,
        0.7131, 0.7134, 0.7135, 0.7136, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137,
        0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7921])
finalReturns:  tensor([1.1096, 1.1880, 1.1467, 1.0220, 0.8385, 0.6128, 0.3567])
----------------------------------------
iter  2  stage  17  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28,
28, 28, 28,
        28, 28, 28, 28, 28, 24,  0])
loss=  tensor(2.0490, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.3838,
4.3838, 4.3838, 4.3838, 4.3838, 4.3838, 4.3838, 4.3838, 4.3838,
        4.3838, 4.3838, 4.3838, 4.3838, 4.3838, 4.3838, 4.3838, 4.3838, 4.3838,
        3.5917, 2.9193, 2.3303, 1.8001, 1.3121, 0.8545, 0.4191]) return=
```

```
173418.33352108797
probs of actions:  tensor([0.9850, 0.9761, 0.9782, 0.9864, 0.9851, 0.9837,
0.9794, 0.9861, 0.9826,
        0.9821, 0.9857, 0.9883, 0.9811, 0.9831, 0.9868, 0.9828, 0.9808, 0.9828,
        0.9931, 0.9881, 0.9930, 0.9938, 0.9869, 0.0494, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4328, 0.5656, 0.6377, 0.6752, 0.6944, 0.7040, 0.7088,
0.7113, 0.7125,
        0.7131, 0.7134, 0.7135, 0.7136, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137,
        0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7345, 0.7744])
finalReturns:  tensor([1.4073, 1.4857, 1.4444, 1.3197, 1.1362, 0.9105, 0.6544,
0.3553])
----------------------------------------
iter  2  stage  16  ep  99999   adversary:  AdversaryModes.myopic
  actions:  tensor([28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28,
28, 28, 28,
        28, 28, 28, 28, 28, 28,  0])
loss=  tensor(0.0740, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.7909,
4.7909, 4.7909, 4.7909, 4.7909, 4.7909, 4.7909, 4.7909, 4.7909,
        4.7909, 4.7909, 4.7909, 4.7909, 4.7909, 4.7909, 4.7909, 4.7909, 3.9988,
        3.3264, 2.7373, 2.2072, 1.7191, 1.2616, 0.8262, 0.4071]) return=
173387.3335190018
probs of actions:  tensor([0.9933, 0.9884, 0.9898, 0.9939, 0.9933, 0.9926,
0.9907, 0.9940, 0.9921,
        0.9919, 0.9937, 0.9949, 0.9914, 0.9925, 0.9942, 0.9922, 0.9930, 0.9938,
        0.9980, 0.9958, 0.9973, 0.9978, 0.9959, 0.9567, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4328, 0.5656, 0.6377, 0.6752, 0.6944, 0.7040, 0.7088,
0.7113, 0.7125,
        0.7131, 0.7134, 0.7135, 0.7136, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137,
        0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7921])
finalReturns:  tensor([1.7108, 1.7892, 1.7479, 1.6233, 1.4397, 1.2141, 0.9579,
0.6796, 0.3850])
----------------------------------------
iter  2  stage  15  ep  52341   adversary:  AdversaryModes.myopic
  actions:  tensor([28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28,
28, 28, 28,
        28, 28, 28, 28, 28, 28,  0])
loss=  tensor(0.0145, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.1891,
5.1891, 5.1891, 5.1891, 5.1891, 5.1891, 5.1891, 5.1891, 5.1891,
        5.1891, 5.1891, 5.1891, 5.1891, 5.1891, 5.1891, 5.1891, 4.3970, 3.7246,
        3.1355, 2.6054, 2.1173, 1.6598, 1.2244, 0.8053, 0.3982]) return=
173387.3335190018
probs of actions:  tensor([0.9987, 0.9975, 0.9980, 0.9989, 0.9988, 0.9986,
0.9983, 0.9989, 0.9985,
        0.9985, 0.9988, 0.9991, 0.9984, 0.9986, 0.9990, 0.9990, 0.9989, 0.9991,
        0.9998, 0.9995, 0.9998, 0.9999, 0.9996, 0.9912, 1.0000],
       grad_fn=<ExpBackward0>)
```

```
rewards:  tensor([0.4328, 0.5656, 0.6377, 0.6752, 0.6944, 0.7040, 0.7088,
0.7113, 0.7125,
        0.7131, 0.7134, 0.7135, 0.7136, 0.7137, 0.7137, 0.7137, 0.7137,
        0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7921])
finalReturns:  tensor([2.0263, 2.1047, 2.0634, 1.9388, 1.7552, 1.5296, 1.2734,
0.9951, 0.7005,
        0.3939])
----------------------------------------
iter  2  stage  14  ep  17   adversary:  AdversaryModes.myopic
  actions:  tensor([28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28,
28, 28, 28,
        28, 28, 28, 28, 28, 28,  0])
loss=  tensor(0.0178, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.5806,
5.5806, 5.5806, 5.5806, 5.5806, 5.5806, 5.5806, 5.5806, 5.5806,
        5.5806, 5.5806, 5.5806, 5.5806, 5.5806, 5.5806, 4.7886, 4.1162, 3.5271,
        2.9969, 2.5089, 2.0513, 1.6160, 1.1969, 0.7898, 0.3916]) return=
173387.3335190018
probs of actions:  tensor([0.9987, 0.9975, 0.9980, 0.9989, 0.9988, 0.9986,
0.9983, 0.9989, 0.9985,
        0.9985, 0.9989, 0.9991, 0.9984, 0.9986, 0.9990, 0.9990, 0.9989, 0.9991,
        0.9998, 0.9995, 0.9998, 0.9999, 0.9996, 0.9913, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4328, 0.5656, 0.6377, 0.6752, 0.6944, 0.7040, 0.7088,
0.7113, 0.7125,
        0.7131, 0.7134, 0.7135, 0.7136, 0.7137, 0.7137, 0.7137, 0.7137,
        0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7921])
finalReturns:  tensor([2.3484, 2.4268, 2.3855, 2.2609, 2.0774, 1.8517, 1.5956,
1.3172, 1.0226,
        0.7160, 0.4005])
----------------------------------------
iter  2  stage  13  ep  4970   adversary:  AdversaryModes.myopic
  actions:  tensor([28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28,
28, 28, 28,
        28, 28, 28, 28, 28, 28,  0])
loss=  tensor(0.0175, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.9673,
5.9673, 5.9673, 5.9673, 5.9673, 5.9673, 5.9673, 5.9673, 5.9673,
        5.9673, 5.9673, 5.9673, 5.9673, 5.9673, 5.1752, 4.5028, 3.9138, 3.3836,
        2.8956, 2.4380, 2.0026, 1.5835, 1.1764, 0.7783, 0.3867]) return=
173387.3335190018
probs of actions:  tensor([0.9990, 0.9979, 0.9984, 0.9991, 0.9990, 0.9989,
0.9986, 0.9991, 0.9988,
        0.9988, 0.9991, 0.9993, 0.9987, 0.9990, 0.9992, 0.9992, 0.9992, 0.9993,
        0.9999, 0.9996, 0.9999, 0.9999, 0.9997, 0.9930, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4328, 0.5656, 0.6377, 0.6752, 0.6944, 0.7040, 0.7088,
0.7113, 0.7125,
        0.7131, 0.7134, 0.7135, 0.7136, 0.7137, 0.7137, 0.7137, 0.7137,
        0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7921])
```

```
finalReturns:  tensor([2.6755, 2.7539, 2.7126, 2.5879, 2.4044, 2.1787, 1.9226,
1.6443, 1.3497,
        1.0431, 0.7275, 0.4054])
----------------------------------------
iter  2  stage  12  ep  205   adversary:  AdversaryModes.myopic
  actions:  tensor([28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28,
28, 28, 28,
        28, 28, 28, 28, 28, 28,  0])
loss=  tensor(0.0208, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.3502,
6.3502, 6.3502, 6.3502, 6.3502, 6.3502, 6.3502, 6.3502, 6.3502,
        6.3502, 6.3502, 6.3502, 6.3502, 5.5582, 4.8858, 4.2968, 3.7666, 3.2786,
        2.8210, 2.3856, 1.9665, 1.5594, 1.1613, 0.7697, 0.3830]) return=
173387.3335190018
probs of actions:  tensor([0.9990, 0.9980, 0.9985, 0.9991, 0.9990, 0.9989,
0.9987, 0.9992, 0.9988,
        0.9988, 0.9991, 0.9993, 0.9990, 0.9990, 0.9993, 0.9992, 0.9992, 0.9993,
        0.9999, 0.9996, 0.9999, 0.9999, 0.9997, 0.9933, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4328, 0.5656, 0.6377, 0.6752, 0.6944, 0.7040, 0.7088,
0.7113, 0.7125,
        0.7131, 0.7134, 0.7135, 0.7136, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137,
        0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7921])
finalReturns:  tensor([3.0062, 3.0846, 3.0433, 2.9186, 2.7351, 2.5094, 2.2533,
1.9750, 1.6804,
        1.3738, 1.0582, 0.7361, 0.4091])
----------------------------------------
iter  2  stage  11  ep  0   adversary:  AdversaryModes.myopic
  actions:  tensor([28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28,
28, 28, 28,
        28, 28, 28, 28, 28, 28,  0])
loss=  tensor(0.0242, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.7303,
6.7303, 6.7303, 6.7303, 6.7303, 6.7303, 6.7303, 6.7303, 6.7303,
        6.7303, 6.7303, 6.7303, 5.9384, 5.2660, 4.6770, 4.1469, 3.6588, 3.2013,
        2.7659, 2.3468, 1.9397, 1.5415, 1.1499, 0.7633, 0.3803]) return=
173387.3335190018
probs of actions:  tensor([0.9990, 0.9980, 0.9985, 0.9991, 0.9990, 0.9989,
0.9987, 0.9992, 0.9988,
        0.9988, 0.9991, 0.9993, 0.9990, 0.9990, 0.9993, 0.9992, 0.9992, 0.9993,
        0.9999, 0.9996, 0.9999, 0.9999, 0.9997, 0.9933, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4328, 0.5656, 0.6377, 0.6752, 0.6944, 0.7040, 0.7088,
0.7113, 0.7125,
        0.7131, 0.7134, 0.7135, 0.7136, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137,
        0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7921])
finalReturns:  tensor([3.3396, 3.4180, 3.3767, 3.2521, 3.0685, 2.8429, 2.5867,
2.3084, 2.0138,
        1.7072, 1.3917, 1.0696, 0.7425, 0.4118])
----------------------------------------
```

```
iter  2  stage  10  ep  0   adversary:  AdversaryModes.myopic
  actions:  tensor([28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28,
28, 28, 28,
        28, 28, 28, 28, 28, 28,  0])
loss=  tensor(0.0286, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.1082,
7.1082, 7.1082, 7.1082, 7.1082, 7.1082, 7.1082, 7.1082, 7.1082, 7.1082,
        7.1082, 7.1082, 6.3164, 5.6442, 5.0552, 4.5251, 4.0370, 3.5795, 3.1441,
        2.7250, 2.3179, 1.9197, 1.5281, 1.1415, 0.7585, 0.3782]) return=
173387.3335190018
probs of actions:  tensor([0.9990, 0.9980, 0.9985, 0.9991, 0.9990, 0.9989,
0.9987, 0.9992, 0.9988,
        0.9988, 0.9991, 0.9993, 0.9990, 0.9990, 0.9993, 0.9992, 0.9992, 0.9993,
        0.9999, 0.9996, 0.9999, 0.9999, 0.9997, 0.9933, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4328, 0.5656, 0.6377, 0.6752, 0.6944, 0.7040, 0.7088,
0.7113, 0.7125,
        0.7131, 0.7134, 0.7135, 0.7136, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137,
        0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7921])
finalReturns:  tensor([3.6751, 3.7535, 3.7122, 3.5875, 3.4040, 3.1783, 2.9222,
2.6439, 2.3493,
        2.0427, 1.7272, 1.4051, 1.0780, 0.7473, 0.4139])
----------------------------------------
iter  2  stage  9  ep  72   adversary:  AdversaryModes.myopic
  actions:  tensor([28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28,
28, 28, 28,
        28, 28, 28, 28, 28, 28,  0])
loss=  tensor(0.0319, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.4843,
7.4843, 7.4843, 7.4843, 7.4843, 7.4843, 7.4843, 7.4843, 7.4843, 7.4843,
        7.4843, 6.6929, 6.0207, 5.4318, 4.9017, 4.4137, 3.9561, 3.5208, 3.1017,
        2.6946, 2.2964, 1.9048, 1.5181, 1.1352, 0.7549, 0.3767]) return=
173387.3335190018
probs of actions:  tensor([0.9990, 0.9981, 0.9985, 0.9992, 0.9991, 0.9989,
0.9988, 0.9992, 0.9989,
        0.9990, 0.9992, 0.9994, 0.9991, 0.9991, 0.9993, 0.9993, 0.9992, 0.9994,
        0.9999, 0.9996, 0.9999, 1.0000, 0.9997, 0.9935, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4328, 0.5656, 0.6377, 0.6752, 0.6944, 0.7040, 0.7088,
0.7113, 0.7125,
        0.7131, 0.7134, 0.7135, 0.7136, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137,
        0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7921])
finalReturns:  tensor([4.0120, 4.0904, 4.0492, 3.9245, 3.7410, 3.5154, 3.2592,
2.9809, 2.6863,
        2.3797, 2.0642, 1.7421, 1.4151, 1.0843, 0.7509, 0.4154])
----------------------------------------
iter  2  stage  8  ep  58   adversary:  AdversaryModes.myopic
  actions:  tensor([28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28,
28, 28, 28,
        28, 28, 28, 28, 28, 28,  0])
```

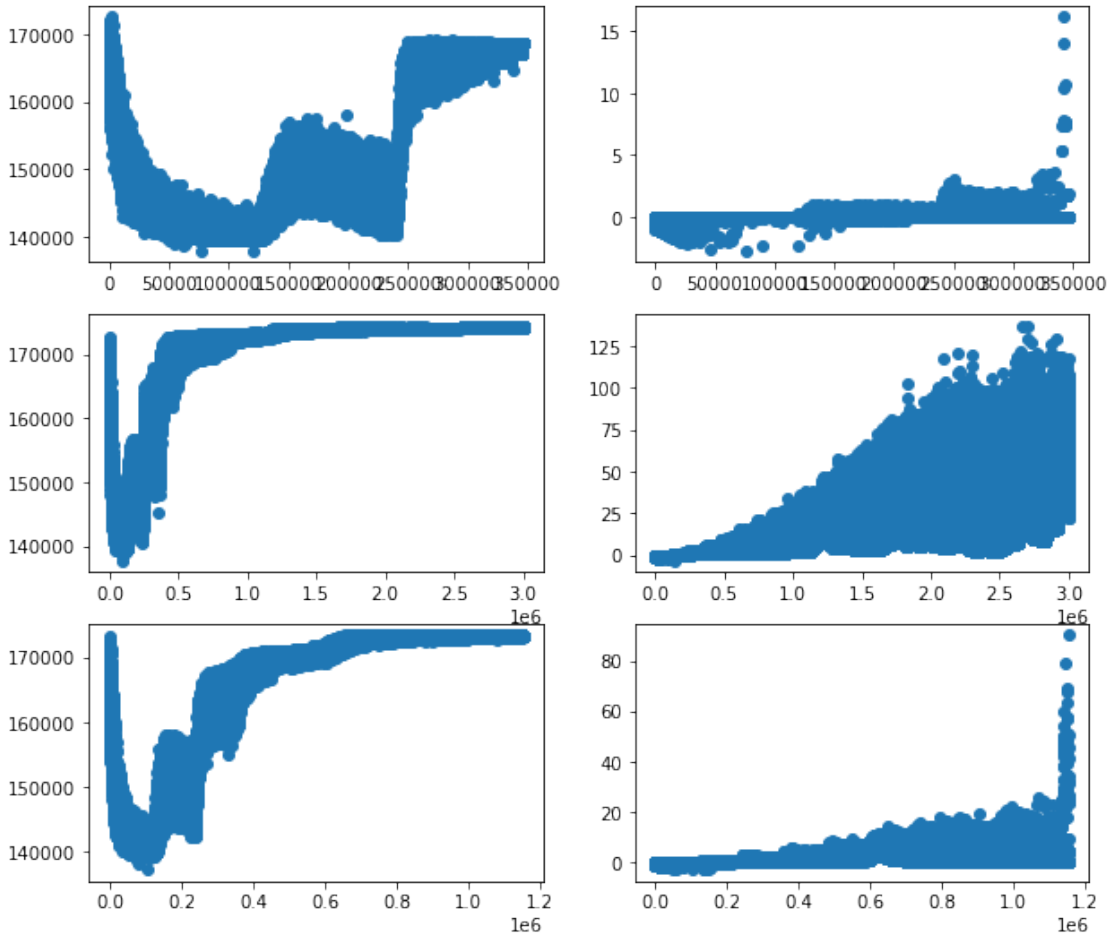
```
loss=  tensor(0.0356, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.8587,
7.8587, 7.8587, 7.8587, 7.8587, 7.8587, 7.8587, 7.8587, 7.8587,
        7.0679, 6.3960, 5.8072, 5.2772, 4.7892, 4.3317, 3.8963, 3.4772, 3.0701,
        2.6719, 2.2803, 1.8937, 1.5107, 1.1304, 0.7522, 0.3755]) return=
173387.3335190018
probs of actions:  tensor([0.9991, 0.9981, 0.9986, 0.9992, 0.9991, 0.9990,
0.9988, 0.9992, 0.9990,
        0.9991, 0.9993, 0.9995, 0.9991, 0.9991, 0.9993, 0.9993, 0.9993, 0.9994,
        0.9999, 0.9997, 0.9999, 1.0000, 0.9997, 0.9937, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4328, 0.5656, 0.6377, 0.6752, 0.6944, 0.7040, 0.7088,
0.7113, 0.7125,
        0.7131, 0.7134, 0.7135, 0.7136, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137,
        0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7921])
finalReturns:  tensor([4.3501, 4.4285, 4.3873, 4.2627, 4.0792, 3.8535, 3.5974,
3.3191, 3.0245,
        2.7179, 2.4024, 2.0803, 1.7532, 1.4225, 1.0891, 0.7536, 0.4166])
----------------------------------------
iter  2  stage  7  ep  0   adversary:  AdversaryModes.myopic
  actions:  tensor([28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28,
28, 28, 28,
        28, 28, 28, 28, 28, 28,  0])
loss=  tensor(0.0404, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.2311,
8.2311, 8.2311, 8.2311, 8.2311, 8.2311, 8.2311, 8.2311, 7.4415,
        6.7702, 6.1817, 5.6517, 5.1638, 4.7063, 4.2709, 3.8519, 3.4448, 3.0466,
        2.6550, 2.2684, 1.8854, 1.5051, 1.1269, 0.7502, 0.3747]) return=
173387.3335190018
probs of actions:  tensor([0.9991, 0.9981, 0.9986, 0.9992, 0.9991, 0.9990,
0.9988, 0.9992, 0.9990,
        0.9991, 0.9993, 0.9995, 0.9991, 0.9991, 0.9993, 0.9993, 0.9993, 0.9994,
        0.9999, 0.9997, 0.9999, 1.0000, 0.9998, 0.9937, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4328, 0.5656, 0.6377, 0.6752, 0.6944, 0.7040, 0.7088,
0.7113, 0.7125,
        0.7131, 0.7134, 0.7135, 0.7136, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137,
        0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7921])
finalReturns:  tensor([4.6890, 4.7674, 4.7262, 4.6016, 4.4182, 4.1925, 3.9364,
3.6581, 3.3635,
        3.0569, 2.7414, 2.4193, 2.0922, 1.7615, 1.4281, 1.0926, 0.7556, 0.4174])
----------------------------------------
iter  2  stage  6  ep  80   adversary:  AdversaryModes.myopic
  actions:  tensor([28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28,
28, 28, 28,
        28, 28, 28, 28, 28, 28,  0])
loss=  tensor(0.0427, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.6006,
8.6006, 8.6006, 8.6006, 8.6006, 8.6006, 8.6006, 7.8134, 7.1432,
        6.5552, 6.0255, 5.5377, 5.0803, 4.6449, 4.2259, 3.8188, 3.4206, 3.0290,
        2.6424, 2.2594, 1.8791, 1.5009, 1.1242, 0.7487, 0.3740]) return=
```

```
173387.3335190018
probs of actions:  tensor([0.9991, 0.9982, 0.9986, 0.9992, 0.9992, 0.9990,
0.9990, 0.9994, 0.9991,
        0.9992, 0.9994, 0.9995, 0.9992, 0.9992, 0.9993, 0.9993, 0.9993, 0.9994,
        0.9999, 0.9997, 0.9999, 1.0000, 0.9998, 0.9941, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4328, 0.5656, 0.6377, 0.6752, 0.6944, 0.7040, 0.7088,
0.7113, 0.7125,
        0.7131, 0.7134, 0.7135, 0.7136, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137,
        0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7921])
finalReturns:  tensor([5.0284, 5.1068, 5.0657, 4.9412, 4.7578, 4.5322, 4.2761,
3.9978, 3.7032,
        3.3966, 3.0811, 2.7589, 2.4319, 2.1012, 1.7678, 1.4323, 1.0953, 0.7571,
        0.4181])
----------------------------------------
iter  2  stage  5  ep  0   adversary:  AdversaryModes.myopic
  actions:  tensor([28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28,
28, 28, 28,
        28, 28, 28, 28, 28, 28,  0])
loss=  tensor(0.0490, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.9651,
8.9651, 8.9651, 8.9651, 8.9651, 8.9651, 8.1827, 7.5147, 6.9278,
        6.3986, 5.9111, 5.4537, 5.0184, 4.5994, 4.1923, 3.7942, 3.4026, 3.0159,
        2.6329, 2.2527, 1.8745, 1.4978, 1.1222, 0.7476, 0.3735]) return=
173387.3335190018
probs of actions:  tensor([0.9991, 0.9982, 0.9986, 0.9992, 0.9992, 0.9990,
0.9990, 0.9994, 0.9991,
        0.9992, 0.9994, 0.9995, 0.9992, 0.9992, 0.9993, 0.9993, 0.9993, 0.9994,
        0.9999, 0.9997, 0.9999, 1.0000, 0.9998, 0.9941, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4328, 0.5656, 0.6377, 0.6752, 0.6944, 0.7040, 0.7088,
0.7113, 0.7125,
        0.7131, 0.7134, 0.7135, 0.7136, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137,
        0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7921])
finalReturns:  tensor([5.3679, 5.4463, 5.4054, 5.2811, 5.0978, 4.8722, 4.6162,
4.3379, 4.0433,
        3.7367, 3.4212, 3.0991, 2.7721, 2.4414, 2.1079, 1.7724, 1.4354, 1.0973,
        0.7582, 0.4186])
----------------------------------------
iter  2  stage  4  ep  0   adversary:  AdversaryModes.myopic
  actions:  tensor([28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28,
28, 28, 28,
        28, 28, 28, 28, 28, 28,  0])
loss=  tensor(0.0550, grad_fn=<NegBackward0>)   ,  base rewards= tensor([9.3201,
9.3201, 9.3201, 9.3201, 9.3201, 8.5474, 7.8839, 7.2990, 6.7709,
        6.2838, 5.8267, 5.3915, 4.9725, 4.5655, 4.1673, 3.7758, 3.3891, 3.0061,
        2.6259, 2.2476, 1.8710, 1.4954, 1.1208, 0.7467, 0.3732]) return=
173387.3335190018
probs of actions:  tensor([0.9991, 0.9982, 0.9986, 0.9992, 0.9992, 0.9990,
```

```
0.9990, 0.9994, 0.9991,
        0.9992, 0.9994, 0.9995, 0.9992, 0.9992, 0.9993, 0.9993, 0.9993, 0.9994,
        0.9999, 0.9997, 0.9999, 1.0000, 0.9998, 0.9941, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4328, 0.5656, 0.6377, 0.6752, 0.6944, 0.7040, 0.7088,
0.7113, 0.7125,
        0.7131, 0.7134, 0.7135, 0.7136, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137,
        0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7921])
finalReturns:  tensor([5.7072, 5.7856, 5.7451, 5.6211, 5.4380, 5.2126, 4.9566,
4.6784, 4.3838,
        4.0772, 3.7617, 3.4396, 3.1126, 2.7819, 2.4484, 2.1130, 1.7759, 1.4378,
        1.0987, 0.7591, 0.4189])
----------------------------------------
iter  2  stage  3  ep  0   adversary:  AdversaryModes.myopic
  actions:  tensor([28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28,
28, 28, 28,
        28, 28, 28, 28, 28, 28,  0])
loss=  tensor(0.0607, grad_fn=<NegBackward0>)   ,  base rewards= tensor([9.6570,
9.6570, 9.6570, 9.6570, 8.9034, 8.2488, 7.6681, 7.1419, 6.6558,
        6.1991, 5.7642, 5.3453, 4.9384, 4.5402, 4.1487, 3.7620, 3.3790, 2.9988,
        2.6206, 2.2439, 1.8683, 1.4937, 1.1196, 0.7461, 0.3729]) return=
173387.3335190018
probs of actions:  tensor([0.9991, 0.9982, 0.9986, 0.9992, 0.9992, 0.9990,
0.9990, 0.9994, 0.9991,
        0.9992, 0.9994, 0.9995, 0.9992, 0.9992, 0.9993, 0.9993, 0.9993, 0.9994,
        0.9999, 0.9997, 0.9999, 1.0000, 0.9998, 0.9941, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4328, 0.5656, 0.6377, 0.6752, 0.6944, 0.7040, 0.7088,
0.7113, 0.7125,
        0.7131, 0.7134, 0.7135, 0.7136, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137,
        0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7921])
finalReturns:  tensor([6.0455, 6.1239, 6.0842, 5.9609, 5.7782, 5.5531, 5.2972,
5.0191, 4.7246,
        4.4180, 4.1025, 3.7804, 3.4534, 3.1227, 2.7892, 2.4537, 2.1167, 1.7786,
        1.4395, 1.0999, 0.7597, 0.4192])
----------------------------------------
iter  2  stage  2  ep  339   adversary:  AdversaryModes.myopic
  actions:  tensor([28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28,
28, 28, 28,
        28, 28, 28, 28, 28, 28,  0])
loss=  tensor(0.0615, grad_fn=<NegBackward0>)   ,  base rewards= tensor([9.9587,
9.9587, 9.9587, 9.2425, 8.6055, 8.0331, 7.5109, 7.0267, 6.5710,
        6.1365, 5.7178, 5.3110, 4.9129, 4.5213, 4.1347, 3.7517, 3.3715, 2.9933,
        2.6166, 2.2411, 1.8664, 1.4924, 1.1188, 0.7456, 0.3727]) return=
173387.3335190018
probs of actions:  tensor([0.9991, 0.9983, 0.9990, 0.9994, 0.9992, 0.9990,
0.9992, 0.9995, 0.9993,
        0.9993, 0.9994, 0.9996, 0.9993, 0.9992, 0.9993, 0.9993, 0.9993, 0.9994,
```

```
        0.9999, 0.9997, 0.9999, 1.0000, 0.9998, 0.9941, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4328, 0.5656, 0.6377, 0.6752, 0.6944, 0.7040, 0.7088,
0.7113, 0.7125,
        0.7131, 0.7134, 0.7135, 0.7136, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137,
        0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7921])
finalReturns:  tensor([6.3816, 6.4600, 6.4218, 6.2998, 6.1181, 5.8934, 5.6379,
5.3599, 5.0655,
        4.7589, 4.4435, 4.1214, 3.7944, 3.4637, 3.1302, 2.7947, 2.4577, 2.1195,
        1.7805, 1.4408, 1.1007, 0.7602, 0.4194])
----------------------------------------
iter  2  stage  1  ep  14789   adversary:  AdversaryModes.myopic
  actions:  tensor([28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28,
28, 28, 28,
        28, 28, 28, 28, 28, 28,  0])
loss=  tensor(0.0368, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([10.1929, 10.1929,  9.5489,  8.9464,  8.3904,  7.8761,  7.3956,  6.9417,
         6.5082,  6.0900,  5.6833,  5.2853,  4.8939,  4.5072,  4.1243,  3.7440,
         3.3658,  2.9891,  2.6136,  2.2389,  1.8649,  1.4914,  1.1182,  0.7453,
         0.3726]) return=  173387.3335190018
probs of actions:  tensor([0.9994, 0.9990, 0.9995, 0.9998, 0.9995, 0.9994,
0.9997, 0.9999, 0.9997,
        1.0000, 0.9997, 1.0000, 0.9996, 0.9996, 0.9996, 0.9996, 0.9996, 0.9998,
        1.0000, 0.9999, 1.0000, 1.0000, 0.9999, 0.9959, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4328, 0.5656, 0.6377, 0.6752, 0.6944, 0.7040, 0.7088,
0.7113, 0.7125,
        0.7131, 0.7134, 0.7135, 0.7136, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137,
        0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7921])
finalReturns:  tensor([6.7130, 6.7914, 6.7562, 6.6369, 6.4569, 6.2333, 5.9784,
5.7007, 5.4064,
        5.1000, 4.7846, 4.4625, 4.1355, 3.8048, 3.4714, 3.1359, 2.7989, 2.4607,
        2.1217, 1.7820, 1.4418, 1.1013, 0.7605, 0.4195])
----------------------------------------
iter  2  stage  0  ep  0   adversary:  AdversaryModes.myopic
  actions:  tensor([28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28,
28, 28, 28,
        28, 28, 28, 28, 28, 28,  0])
loss=  tensor(0.0418, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([10.3041,  9.7929,  9.2563,  8.7325,  8.2337,  7.7608,  7.3106,  6.8789,
         6.4615,  6.0553,  5.6576,  5.2662,  4.8796,  4.4967,  4.1164,  3.7383,
         3.3616,  2.9860,  2.6114,  2.2374,  1.8638,  1.4906,  1.1177,  0.7450,
         0.3724]) return=  173387.3335190018
probs of actions:  tensor([0.9994, 0.9990, 0.9995, 0.9998, 0.9995, 0.9994,
0.9997, 0.9999, 0.9997,
        1.0000, 0.9997, 1.0000, 0.9996, 0.9996, 0.9996, 0.9996, 0.9996, 0.9998,
        1.0000, 0.9999, 1.0000, 1.0000, 0.9999, 0.9959, 1.0000],
       grad_fn=<ExpBackward0>)
```

```
rewards:  tensor([0.4328, 0.5656, 0.6377, 0.6752, 0.6944, 0.7040, 0.7088,
0.7113, 0.7125,
        0.7131, 0.7134, 0.7135, 0.7136, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137,
        0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7137, 0.7921])
finalReturns:  tensor([7.0346, 7.1130, 7.0840, 6.9701, 6.7936, 6.5721, 6.3183,
6.0413, 5.7473,
        5.4411, 5.1257, 4.8037, 4.4767, 4.1460, 3.8126, 3.4771, 3.1401, 2.8019,
        2.4629, 2.1232, 1.7831, 1.4426, 1.1018, 0.7608, 0.4197])
0,[1e-05,1][1, 10000, 1, 1],1682452608 saved
[1152887, 'tensor([1., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.])',
173387.3335190018, 58923.33323528369, 0.041793305426836014, 1e-05, 1, 0,
'tensor([28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 28,
28,\n        28, 28, 28, 28, 28, 28,  0])', '[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.\n 1.]', '0,[1e-05,1][1, 10000, 1,
1],1682452608', 25, 50, 173387.33351900178, 222740.53173358264,
93054.95743178323, 131763.112, 128763.66666666667, 68687.38784334851,
68702.99934306633, 84336.85885052304, 84336.85885052304, 107493.35770590571,
68702.99934306633, 84336.85885052304]
```



```
policy reset
----------------------------------------
iter  0  stage  24  ep  99999   adversary:  AdversaryModes.constant_132
  actions:  tensor([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 1,
        0])
loss=  tensor(-0., grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.5662,
0.5662, 0.5662, 0.5662, 0.5662, 0.5662, 0.5662, 0.5662, 0.5662,
        0.5662, 0.5662, 0.5662, 0.5662, 0.5662, 0.5662, 0.5662, 0.5662, 0.5662,
        0.5662, 0.5662, 0.5662, 0.5662, 0.5662, 0.5662, 0.5662]) return=
138591.14106947038
probs of actions:  tensor([0.9412, 0.9369, 0.9014, 0.9244, 0.9103, 0.9282,
0.9200, 0.9465, 0.9221,
        0.9313, 0.9320, 0.9221, 0.9166, 0.9288, 0.9280, 0.9255, 0.9292, 0.9279,
        0.9284, 0.9441, 0.9064, 0.9456, 0.9071, 0.0328, 0.9943],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5112, 0.5238, 0.5334, 0.5406, 0.5460, 0.5501, 0.5532,
0.5555, 0.5573,
        0.5586, 0.5595, 0.5603, 0.5608, 0.5613, 0.5616, 0.5618, 0.5620, 0.5621,
        0.5622, 0.5623, 0.5623, 0.5624, 0.5624, 0.5623, 0.5662])
finalReturns:  tensor([0.])
----------------------------------------
iter  0  stage  23  ep  99999   adversary:  AdversaryModes.constant_132
  actions:  tensor([ 1,  0, 23, 15,  2,  0,  0,  0,  0,  0,  0,  0,  0, 15,  0,
14,  1,  0,
        13,  1,  1,  3,  0, 15,  0])
loss=  tensor(0.0938, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.1960,
1.1960, 1.1960, 1.1960, 1.1960, 1.1960, 1.1960, 1.1960, 1.1960,
        1.1960, 1.1960, 1.1960, 1.1960, 1.1960, 1.1960, 1.1960, 1.1960, 1.1960,
        1.1960, 1.1960, 1.1960, 1.1960, 1.1960, 1.1960, 0.5929]) return=
150337.40439247817
probs of actions:  tensor([0.0699, 0.5186, 0.0031, 0.0702, 0.0520, 0.4937,
0.4496, 0.6094, 0.4886,
        0.5172, 0.5479, 0.4457, 0.5186, 0.0648, 0.4107, 0.0057, 0.0700, 0.4843,
        0.0029, 0.0619, 0.0670, 0.0427, 0.5052, 0.0788, 0.9895],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5111, 0.5274, 0.4832, 0.6082, 0.6730, 0.6528, 0.6296,
0.6125, 0.5998,
        0.5903, 0.5833, 0.5781, 0.5742, 0.5487, 0.6270, 0.5910, 0.6536, 0.6343,
        0.5990, 0.6537, 0.6342, 0.6190, 0.6170, 0.5807, 0.6520])
finalReturns:  tensor([0.0367, 0.0592])
----------------------------------------
iter  0  stage  22  ep  99999   adversary:  AdversaryModes.constant_132
  actions:  tensor([ 0, 22, 22, 22, 18, 23, 22, 22, 27, 22, 22, 21, 22, 26, 22,
0, 17,  0,
        22, 22, 16,  0, 28, 22,  0])
loss=  tensor(0.8564, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.1470,
2.1470, 2.1470, 2.1470, 2.1470, 2.1470, 2.1470, 2.1470, 2.1470,
```

```
        2.1470, 2.1470, 2.1470, 2.1470, 2.1470, 2.1470, 2.1470, 2.1470, 2.1470,
        2.1470, 2.1470, 2.1470, 2.1470, 2.1470, 1.3829, 0.6721]) return=
193459.92049804344
probs of actions:  tensor([0.2512, 0.3682, 0.3488, 0.2969, 0.0235, 0.0110,
0.3700, 0.3502, 0.0278,
        0.3349, 0.3164, 0.0416, 0.3236, 0.0325, 0.4630, 0.1016, 0.0405, 0.1303,
        0.3723, 0.4079, 0.0380, 0.3571, 0.0170, 0.6691, 0.9944],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5112, 0.4754, 0.5683, 0.6430, 0.7178, 0.7252, 0.7735,
0.8025, 0.8000,
        0.8650, 0.8719, 0.8813, 0.8760, 0.8609, 0.9026, 0.9485, 0.8137, 0.8435,
        0.7196, 0.7612, 0.8159, 0.8381, 0.6857, 0.7853, 0.8599])
finalReturns:  tensor([0.1839, 0.2623, 0.1878])
----------------------------------------
iter  0  stage  21  ep  99999   adversary:  AdversaryModes.constant_132
  actions:  tensor([30, 30, 22, 19, 30, 30, 30, 27, 20, 27, 30, 28, 30, 30, 26,
30, 26, 22,
        30, 24, 30, 24, 17, 27,  0])
loss=  tensor(4.5377, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.5209,
3.5209, 3.5209, 3.5209, 3.5209, 3.5209, 3.5209, 3.5209, 3.5209,
        3.5209, 3.5209, 3.5209, 3.5209, 3.5209, 3.5209, 3.5209, 3.5209, 3.5209,
        3.5209, 3.5209, 3.5209, 3.5209, 2.4752, 1.5642, 0.7482]) return=
219456.57013693356
probs of actions:  tensor([0.3420, 0.3385, 0.2044, 0.0172, 0.4148, 0.3398,
0.4331, 0.0786, 0.0263,
        0.0908, 0.4014, 0.0236, 0.3730, 0.3527, 0.1123, 0.3745, 0.1065, 0.2665,
        0.3297, 0.0252, 0.3049, 0.0192, 0.0083, 0.0535, 0.9814],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4212, 0.5480, 0.6939, 0.7536, 0.7227, 0.7910, 0.8441,
0.9020, 0.9510,
        0.8956, 0.8963, 0.9363, 0.9362, 0.9550, 0.9916, 0.9594, 0.9949, 1.0034,
        0.9335, 0.9854, 0.9371, 0.9881, 1.0002, 0.9088, 0.9963])
finalReturns:  tensor([0.3725, 0.4301, 0.3409, 0.2480])
----------------------------------------
iter  0  stage  20  ep  99999   adversary:  AdversaryModes.constant_132
  actions:  tensor([30, 30, 30, 30, 30, 30, 30, 30, 27, 30, 30, 30, 30, 30, 30,
30, 30, 30,
        30, 30, 30, 30, 30, 30,  0])
loss=  tensor(0.1581, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.3714,
4.3714, 4.3714, 4.3714, 4.3714, 4.3714, 4.3714, 4.3714, 4.3714,
        4.3714, 4.3714, 4.3714, 4.3714, 4.3714, 4.3714, 4.3714, 4.3714, 4.3714,
        4.3714, 4.3714, 4.3714, 3.2718, 2.3232, 1.4805, 0.7132]) return=
228028.54094405496
probs of actions:  tensor([0.9125, 0.9169, 0.9037, 0.9342, 0.9368, 0.8968,
0.9510, 0.9254, 0.0124,
        0.9645, 0.9396, 0.9534, 0.9234, 0.9067, 0.9477, 0.9371, 0.9293, 0.8873,
        0.8996, 0.9431, 0.9692, 0.9912, 0.9642, 0.8582, 0.9988],
       grad_fn=<ExpBackward0>)
```

```
rewards:  tensor([0.4212, 0.5480, 0.6523, 0.7357, 0.8011, 0.8519, 0.8908,
0.9206, 0.9603,
        0.9450, 0.9617, 0.9743, 0.9838, 0.9909, 0.9963, 1.0003, 1.0034, 1.0056,
        1.0074, 1.0086, 1.0096, 1.0103, 1.0109, 1.0113, 1.1016])
finalReturns:  tensor([0.7723, 0.8623, 0.8006, 0.6324, 0.3884])
---------------------------------------
iter  0  stage  19  ep  99999   adversary:  AdversaryModes.constant_132
  actions:  tensor([30, 30, 28, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30,
30, 30, 30,
        30, 30, 30, 30, 30, 30,  0])
loss=  tensor(0.0536, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.0447,
5.0447, 5.0447, 5.0447, 5.0447, 5.0447, 5.0447, 5.0447, 5.0447, 5.0447,
        5.0447, 5.0447, 5.0447, 5.0447, 5.0447, 5.0447, 5.0447, 5.0447, 5.0447,
        5.0447, 5.0447, 3.9452, 2.9968, 2.1542, 1.3869, 0.6738]) return=
228200.9420715731
probs of actions:  tensor([0.9825, 0.9829, 0.0017, 0.9874, 0.9864, 0.9742,
0.9898, 0.9839, 0.9872,
        0.9938, 0.9878, 0.9901, 0.9837, 0.9793, 0.9897, 0.9861, 0.9859, 0.9749,
        0.9783, 0.9932, 0.9963, 0.9995, 0.9941, 0.9526, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4212, 0.5480, 0.6639, 0.7266, 0.7941, 0.8464, 0.8867,
0.9174, 0.9408,
        0.9585, 0.9719, 0.9820, 0.9895, 0.9953, 0.9996, 1.0028, 1.0052, 1.0070,
        1.0084, 1.0094, 1.0102, 1.0108, 1.0112, 1.0115, 1.1018])
finalReturns:  tensor([1.1102, 1.2002, 1.1385, 0.9703, 0.7264, 0.4280])
---------------------------------------
iter  0  stage  18  ep  86927   adversary:  AdversaryModes.constant_132
  actions:  tensor([30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30,
30, 30, 30,
        30, 30, 30, 30, 30, 30,  0])
loss=  tensor(0.0098, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.6869,
5.6869, 5.6869, 5.6869, 5.6869, 5.6869, 5.6869, 5.6869, 5.6869, 5.6869,
        5.6869, 5.6869, 5.6869, 5.6869, 5.6869, 5.6869, 5.6869, 5.6869, 5.6869,
        5.6869, 4.5884, 3.6405, 2.7984, 2.0314, 1.3185, 0.6449]) return=
228471.31797735966
probs of actions:  tensor([0.9982, 0.9981, 0.9972, 0.9989, 0.9985, 0.9973,
0.9989, 0.9984, 0.9987,
        0.9994, 0.9987, 0.9990, 0.9983, 0.9981, 0.9990, 0.9985, 0.9986, 0.9974,
        0.9990, 0.9997, 1.0000, 1.0000, 0.9995, 0.9908, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4212, 0.5480, 0.6523, 0.7357, 0.8011, 0.8519, 0.8908,
0.9206, 0.9432,
        0.9603, 0.9732, 0.9830, 0.9903, 0.9959, 1.0000, 1.0031, 1.0055, 1.0072,
        1.0085, 1.0095, 1.0103, 1.0108, 1.0112, 1.0116, 1.1018])
finalReturns:  tensor([1.4768, 1.5668, 1.5052, 1.3371, 1.0932, 0.7948, 0.4569])
---------------------------------------
iter  0  stage  17  ep  20360   adversary:  AdversaryModes.constant_132
  actions:  tensor([30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30,
```

```
30, 30, 30,
        30, 30, 30, 30, 30, 30,  0])
loss=  tensor(0.0088, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.3064,
6.3064, 6.3064, 6.3064, 6.3064, 6.3064, 6.3064, 6.3064, 6.3064,
        6.3064, 6.3064, 6.3064, 6.3064, 6.3064, 6.3064, 6.3064, 6.3064, 6.3064,
        5.2092, 4.2623, 3.4208, 2.6543, 1.9417, 1.2684, 0.6236]) return=
228471.31797735966
probs of actions:  tensor([0.9991, 0.9989, 0.9985, 0.9994, 0.9992, 0.9986,
0.9994, 0.9992, 0.9993,
        0.9997, 0.9993, 0.9995, 0.9991, 0.9990, 0.9995, 0.9992, 0.9992, 0.9990,
        0.9998, 1.0000, 1.0000, 1.0000, 0.9999, 0.9926, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4212, 0.5480, 0.6523, 0.7357, 0.8011, 0.8519, 0.8908,
0.9206, 0.9432,
        0.9603, 0.9732, 0.9830, 0.9903, 0.9959, 1.0000, 1.0031, 1.0055, 1.0072,
        1.0085, 1.0095, 1.0103, 1.0108, 1.0112, 1.0116, 1.1018])
finalReturns:  tensor([1.8645, 1.9545, 1.8929, 1.7249, 1.4811, 1.1829, 0.8450,
0.4782])
----------------------------------------
iter  0  stage  16  ep  0   adversary:  AdversaryModes.constant_132
  actions:  tensor([30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30,
30, 30, 30,
        30, 30, 30, 30, 30, 30,  0])
loss=  tensor(0.0114, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.9087,
6.9087, 6.9087, 6.9087, 6.9087, 6.9087, 6.9087, 6.9087, 6.9087,
        6.9087, 6.9087, 6.9087, 6.9087, 6.9087, 6.9087, 6.9087, 6.9087, 5.8133,
        4.8675, 4.0269, 3.2610, 2.5489, 1.8759, 1.2314, 0.6079]) return=
228471.31797735966
probs of actions:  tensor([0.9991, 0.9989, 0.9985, 0.9994, 0.9992, 0.9986,
0.9994, 0.9992, 0.9993,
        0.9997, 0.9993, 0.9995, 0.9991, 0.9990, 0.9995, 0.9992, 0.9992, 0.9990,
        0.9998, 1.0000, 1.0000, 1.0000, 0.9999, 0.9926, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4212, 0.5480, 0.6523, 0.7357, 0.8011, 0.8519, 0.8908,
0.9206, 0.9432,
        0.9603, 0.9732, 0.9830, 0.9903, 0.9959, 1.0000, 1.0031, 1.0055, 1.0072,
        1.0085, 1.0095, 1.0103, 1.0108, 1.0112, 1.0116, 1.1018])
finalReturns:  tensor([2.2677, 2.3577, 2.2962, 2.1283, 1.8847, 1.5865, 1.2487,
0.8820, 0.4939])
----------------------------------------
iter  0  stage  15  ep  0   adversary:  AdversaryModes.constant_132
  actions:  tensor([30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30,
30, 30, 30,
        30, 30, 30, 30, 30, 30,  0])
loss=  tensor(0.0146, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.4973,
7.4973, 7.4973, 7.4973, 7.4973, 7.4973, 7.4973, 7.4973, 7.4973,
        7.4973, 7.4973, 7.4973, 7.4973, 7.4973, 7.4973, 7.4973, 6.4042, 5.4601,
        4.6206, 3.8555, 3.1440, 2.4714, 1.8273, 1.2040, 0.5963]) return=
```

```
228471.31797735966
probs of actions:  tensor([0.9991, 0.9989, 0.9985, 0.9994, 0.9992, 0.9986,
0.9994, 0.9992, 0.9993,
        0.9997, 0.9993, 0.9995, 0.9991, 0.9990, 0.9995, 0.9992, 0.9992, 0.9990,
        0.9998, 1.0000, 1.0000, 1.0000, 0.9999, 0.9926, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4212, 0.5480, 0.6523, 0.7357, 0.8011, 0.8519, 0.8908,
0.9206, 0.9432,
        0.9603, 0.9732, 0.9830, 0.9903, 0.9959, 1.0000, 1.0031, 1.0055, 1.0072,
        1.0085, 1.0095, 1.0103, 1.0108, 1.0112, 1.0116, 1.1018])
finalReturns:  tensor([2.6823, 2.7723, 2.7109, 2.5432, 2.2997, 2.0017, 1.6640,
1.2973, 0.9093,
        0.5055])
----------------------------------------
iter  0  stage  14  ep  0   adversary:  AdversaryModes.constant_132
  actions:  tensor([30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30,
30, 30, 30,
        30, 30, 30, 30, 30, 30,  0])
loss=  tensor(0.0173, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.0744,
8.0744, 8.0744, 8.0744, 8.0744, 8.0744, 8.0744, 8.0744, 8.0744,
        8.0744, 8.0744, 8.0744, 8.0744, 8.0744, 8.0744, 6.9844, 6.0425, 5.2046,
        4.4406, 3.7299, 3.0579, 2.4141, 1.7912, 1.1837, 0.5876]) return=
228471.31797735966
probs of actions:  tensor([0.9991, 0.9989, 0.9985, 0.9994, 0.9992, 0.9986,
0.9994, 0.9992, 0.9993,
        0.9997, 0.9993, 0.9995, 0.9991, 0.9990, 0.9995, 0.9992, 0.9992, 0.9990,
        0.9998, 1.0000, 1.0000, 1.0000, 0.9999, 0.9926, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4212, 0.5480, 0.6523, 0.7357, 0.8011, 0.8519, 0.8908,
0.9206, 0.9432,
        0.9603, 0.9732, 0.9830, 0.9903, 0.9959, 1.0000, 1.0031, 1.0055, 1.0072,
        1.0085, 1.0095, 1.0103, 1.0108, 1.0112, 1.0116, 1.1018])
finalReturns:  tensor([3.1051, 3.1951, 3.1339, 2.9664, 2.7231, 2.4253, 2.0878,
1.7213, 1.3334,
        0.9296, 0.5142])
----------------------------------------
iter  0  stage  13  ep  2683   adversary:  AdversaryModes.constant_132
  actions:  tensor([30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30,
30, 30, 30,
        30, 30, 30, 30, 30, 30,  0])
loss=  tensor(0.0196, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.6415,
8.6415, 8.6415, 8.6415, 8.6415, 8.6415, 8.6415, 8.6415, 8.6415,
        8.6415, 8.6415, 8.6415, 8.6415, 8.6415, 7.5556, 6.6166, 5.7807, 5.0182,
        4.3086, 3.6373, 2.9942, 2.3717, 1.7645, 1.1686, 0.5812]) return=
228471.31797735966
probs of actions:  tensor([0.9991, 0.9989, 0.9986, 0.9994, 0.9992, 0.9986,
0.9994, 0.9992, 0.9993,
        0.9997, 0.9993, 0.9995, 0.9991, 0.9990, 0.9997, 0.9993, 0.9995, 0.9990,
```

```
        0.9998, 1.0000, 1.0000, 1.0000, 0.9999, 0.9926, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4212, 0.5480, 0.6523, 0.7357, 0.8011, 0.8519, 0.8908,
0.9206, 0.9432,
        0.9603, 0.9732, 0.9830, 0.9903, 0.9959, 1.0000, 1.0031, 1.0055, 1.0072,
        1.0085, 1.0095, 1.0103, 1.0108, 1.0112, 1.0116, 1.1018])
finalReturns:  tensor([3.5339, 3.6239, 3.5629, 3.3957, 3.1527, 2.8552, 2.5179,
2.1515, 1.7637,
        1.3601, 0.9448, 0.5206])
----------------------------------------
iter  0  stage  12  ep  0   adversary:  AdversaryModes.constant_132
  actions:  tensor([30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30,
30, 30, 30,
        30, 30, 30, 30, 30, 30,  0])
loss=  tensor(0.0243, grad_fn=<NegBackward0>)   ,  base rewards= tensor([9.1989,
9.1989, 9.1989, 9.1989, 9.1989, 9.1989, 9.1989, 9.1989, 9.1989,
        9.1989, 9.1989, 9.1989, 9.1989, 8.1186, 7.1834, 6.3502, 5.5897, 4.8815,
        4.2112, 3.5688, 2.9469, 2.3401, 1.7445, 1.1573, 0.5763]) return=
228471.31797735966
probs of actions:  tensor([0.9991, 0.9989, 0.9986, 0.9994, 0.9992, 0.9986,
0.9994, 0.9992, 0.9993,
        0.9997, 0.9993, 0.9995, 0.9991, 0.9990, 0.9997, 0.9993, 0.9995, 0.9990,
        0.9998, 1.0000, 1.0000, 1.0000, 0.9999, 0.9926, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4212, 0.5480, 0.6523, 0.7357, 0.8011, 0.8519, 0.8908,
0.9206, 0.9432,
        0.9603, 0.9732, 0.9830, 0.9903, 0.9959, 1.0000, 1.0031, 1.0055, 1.0072,
        1.0085, 1.0095, 1.0103, 1.0108, 1.0112, 1.0116, 1.1018])
finalReturns:  tensor([3.9668, 4.0568, 3.9962, 3.8293, 3.5867, 3.2895, 2.9525,
2.5864, 2.1988,
        1.7953, 1.3801, 0.9560, 0.5255])
----------------------------------------
iter  0  stage  11  ep  0   adversary:  AdversaryModes.constant_132
  actions:  tensor([30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30,
30, 30, 30,
        30, 30, 30, 30, 30, 30,  0])
loss=  tensor(0.0279, grad_fn=<NegBackward0>)   ,  base rewards= tensor([9.7463,
9.7463, 9.7463, 9.7463, 9.7463, 9.7463, 9.7463, 9.7463, 9.7463,
        9.7463, 9.7463, 9.7463, 8.6733, 7.7432, 6.9137, 6.1557, 5.4494, 4.7805,
        4.1391, 3.5180, 2.9117, 2.3166, 1.7297, 1.1489, 0.5727]) return=
228471.31797735966
probs of actions:  tensor([0.9991, 0.9989, 0.9986, 0.9994, 0.9992, 0.9986,
0.9994, 0.9992, 0.9993,
        0.9997, 0.9993, 0.9995, 0.9991, 0.9990, 0.9997, 0.9993, 0.9995, 0.9990,
        0.9998, 1.0000, 1.0000, 1.0000, 0.9999, 0.9926, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4212, 0.5480, 0.6523, 0.7357, 0.8011, 0.8519, 0.8908,
0.9206, 0.9432,
```

```
        0.9603, 0.9732, 0.9830, 0.9903, 0.9959, 1.0000, 1.0031, 1.0055, 1.0072,
        1.0085, 1.0095, 1.0103, 1.0108, 1.0112, 1.0116, 1.1018])
finalReturns:  tensor([4.4025, 4.4925, 4.4322, 4.2659, 4.0238, 3.7270, 3.3904,
3.0246, 2.6373,
        2.2340, 1.8189, 1.3949, 0.9645, 0.5291])
---------------------------------------
iter  0  stage  10  ep  0   adversary:  AdversaryModes.constant_132
  actions:  tensor([30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30,
30, 30, 30,
        30, 30, 30, 30, 30, 30,  0])
loss=  tensor(0.0325, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([10.2824, 10.2824, 10.2824, 10.2824, 10.2824, 10.2824, 10.2824, 10.2824,
        10.2824, 10.2824, 10.2824,  9.2192,  8.2959,  7.4712,  6.7167,  6.0129,
         5.3459,  4.7058,  4.0856,  3.4801,  2.8855,  2.2990,  1.7185,  1.1426,
         0.5700]) return=  228471.31797735966
probs of actions:  tensor([0.9991, 0.9989, 0.9986, 0.9994, 0.9992, 0.9986,
0.9994, 0.9992, 0.9993,
        0.9997, 0.9993, 0.9995, 0.9991, 0.9990, 0.9997, 0.9993, 0.9995, 0.9990,
        0.9998, 1.0000, 1.0000, 1.0000, 0.9999, 0.9926, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4212, 0.5480, 0.6523, 0.7357, 0.8011, 0.8519, 0.8908,
0.9206, 0.9432,
        0.9603, 0.9732, 0.9830, 0.9903, 0.9959, 1.0000, 1.0031, 1.0055, 1.0072,
        1.0085, 1.0095, 1.0103, 1.0108, 1.0112, 1.0116, 1.1018])
finalReturns:  tensor([4.8396, 4.9296, 4.8698, 4.7042, 4.4628, 4.1667, 3.8306,
3.4652, 3.0782,
        2.6751, 2.2602, 1.8364, 1.4061, 0.9708, 0.5318])
---------------------------------------
iter  0  stage  9  ep  0   adversary:  AdversaryModes.constant_132
  actions:  tensor([30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30,
30, 30, 30,
        30, 30, 30, 30, 30, 30,  0])
loss=  tensor(0.0354, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([10.8052, 10.8052, 10.8052, 10.8052, 10.8052, 10.8052, 10.8052, 10.8052,
        10.8052, 10.8052,  9.7549,  8.8407,  8.0224,  7.2725,  6.5720,  5.9074,
         5.2692,  4.6503,  4.0457,  3.4519,  2.8659,  2.2859,  1.7102,  1.1379,
         0.5680]) return=  228471.31797735966
probs of actions:  tensor([0.9991, 0.9989, 0.9986, 0.9994, 0.9992, 0.9986,
0.9994, 0.9992, 0.9993,
        0.9997, 0.9993, 0.9995, 0.9991, 0.9990, 0.9997, 0.9993, 0.9995, 0.9990,
        0.9998, 1.0000, 1.0000, 1.0000, 0.9999, 0.9926, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4212, 0.5480, 0.6523, 0.7357, 0.8011, 0.8519, 0.8908,
0.9206, 0.9432,
        0.9603, 0.9732, 0.9830, 0.9903, 0.9959, 1.0000, 1.0031, 1.0055, 1.0072,
        1.0085, 1.0095, 1.0103, 1.0108, 1.0112, 1.0116, 1.1018])
finalReturns:  tensor([5.2771, 5.3671, 5.3080, 5.1433, 4.9029, 4.6075, 4.2721,
3.9073, 3.5207,
```

```
        3.1180, 2.7034, 2.2798, 1.8496, 1.4144, 0.9755, 0.5338])
---------------------------------------
iter  0  stage  8  ep  0   adversary:  AdversaryModes.constant_132
  actions:  tensor([30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30,
30, 30, 30,
        30, 30, 30, 30, 30, 30,  0])
loss=  tensor(0.0407, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([11.3115, 11.3115, 11.3115, 11.3115, 11.3115, 11.3115, 11.3115, 11.3115,
        11.3115, 10.2783,  9.3761,  8.5663,  7.8226,  7.1265,  6.4652,  5.8293,
         5.2122,  4.6089,  4.0160,  3.4308,  2.8513,  2.2760,  1.7040,  1.1344,
         0.5665]) return=  228471.31797735966
probs of actions:  tensor([0.9991, 0.9989, 0.9986, 0.9994, 0.9992, 0.9986,
0.9994, 0.9992, 0.9993,
        0.9997, 0.9993, 0.9995, 0.9991, 0.9990, 0.9997, 0.9993, 0.9995, 0.9990,
        0.9998, 1.0000, 1.0000, 1.0000, 0.9999, 0.9926, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4212, 0.5480, 0.6523, 0.7357, 0.8011, 0.8519, 0.8908,
0.9206, 0.9432,
        0.9603, 0.9732, 0.9830, 0.9903, 0.9959, 1.0000, 1.0031, 1.0055, 1.0072,
        1.0085, 1.0095, 1.0103, 1.0108, 1.0112, 1.0116, 1.1018])
finalReturns:  tensor([5.7139, 5.8039, 5.7458, 5.5824, 5.3432, 5.0489, 4.7144,
4.3503, 3.9643,
        3.5620, 3.1477, 2.7244, 2.2944, 1.8594, 1.4206, 0.9790, 0.5353])
---------------------------------------
iter  0  stage  7  ep  0   adversary:  AdversaryModes.constant_132
  actions:  tensor([30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30,
30, 30, 30,
        30, 30, 30, 30, 30, 30,  0])
loss=  tensor(0.0468, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([11.7970, 11.7970, 11.7970, 11.7970, 11.7970, 11.7970, 11.7970, 11.7970,
        10.7864,  9.9001,  9.1016,  8.3659,  7.6757,  7.0187,  6.3860,  5.7712,
         5.1697,  4.5781,  3.9938,  3.4150,  2.8403,  2.2687,  1.6994,  1.1317,
         0.5654]) return=  228471.31797735966
probs of actions:  tensor([0.9991, 0.9989, 0.9986, 0.9994, 0.9992, 0.9986,
0.9994, 0.9992, 0.9993,
        0.9997, 0.9993, 0.9995, 0.9991, 0.9990, 0.9997, 0.9993, 0.9995, 0.9990,
        0.9998, 1.0000, 1.0000, 1.0000, 0.9999, 0.9926, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4212, 0.5480, 0.6523, 0.7357, 0.8011, 0.8519, 0.8908,
0.9206, 0.9432,
        0.9603, 0.9732, 0.9830, 0.9903, 0.9959, 1.0000, 1.0031, 1.0055, 1.0072,
        1.0085, 1.0095, 1.0103, 1.0108, 1.0112, 1.0116, 1.1018])
finalReturns:  tensor([6.1490, 6.2390, 6.1822, 6.0204, 5.7828, 5.4900, 5.1567,
4.7936, 4.4084,
        4.0067, 3.5929, 3.1699, 2.7402, 2.3054, 1.8667, 1.4252, 0.9816, 0.5364])
---------------------------------------
iter  0  stage  6  ep  0   adversary:  AdversaryModes.constant_132
  actions:  tensor([30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30,
```

```
30, 30, 30,
        30, 30, 30, 30, 30, 30,  0])
loss=  tensor(0.0518, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([12.2557, 12.2557, 12.2557, 12.2557, 12.2557, 12.2557, 12.2557, 11.2749,
        10.4095,  9.6259,  8.9010,  8.2186,  7.5673,  6.9388,  6.3271,  5.7279,
         5.1380,  4.5550,  3.9772,  3.4032,  2.8321,  2.2632,  1.6959,  1.1297,
         0.5645]) return=  228471.31797735966
probs of actions:  tensor([0.9991, 0.9989, 0.9986, 0.9994, 0.9992, 0.9986,
0.9994, 0.9992, 0.9993,
        0.9997, 0.9993, 0.9995, 0.9991, 0.9990, 0.9997, 0.9993, 0.9995, 0.9990,
        0.9998, 1.0000, 1.0000, 1.0000, 0.9999, 0.9926, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4212, 0.5480, 0.6523, 0.7357, 0.8011, 0.8519, 0.8908,
0.9206, 0.9432,
        0.9603, 0.9732, 0.9830, 0.9903, 0.9959, 1.0000, 1.0031, 1.0055, 1.0072,
        1.0085, 1.0095, 1.0103, 1.0108, 1.0112, 1.0116, 1.1018])
finalReturns:  tensor([6.5812, 6.6712, 6.6160, 6.4564, 6.2210, 5.9301, 5.5985,
5.2366, 4.8524,
        4.4516, 4.0384, 3.6159, 3.1865, 2.7520, 2.3136, 1.8722, 1.4287, 0.9836,
        0.5373])
----------------------------------------
iter  0  stage  5  ep  280   adversary:  AdversaryModes.constant_132
  actions:  tensor([30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30,
30, 30, 30,
        30, 30, 30, 30, 30, 30,  0])
loss=  tensor(0.0541, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([12.6800, 12.6800, 12.6800, 12.6800, 12.6800, 12.6800, 11.7381, 10.9002,
        10.1362,  9.4255,  8.7535,  8.1098,  7.4869,  6.8794,  6.2833,  5.6957,
         5.1144,  4.5378,  3.9648,  3.3944,  2.8260,  2.2591,  1.6933,  1.1283,
         0.5639]) return=  228471.31797735966
probs of actions:  tensor([0.9991, 0.9989, 0.9985, 0.9994, 0.9992, 0.9990,
0.9994, 0.9994, 0.9995,
        0.9998, 0.9995, 0.9996, 0.9993, 0.9990, 0.9997, 0.9993, 0.9995, 0.9990,
        0.9998, 1.0000, 1.0000, 1.0000, 0.9999, 0.9921, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4212, 0.5480, 0.6523, 0.7357, 0.8011, 0.8519, 0.8908,
0.9206, 0.9432,
        0.9603, 0.9732, 0.9830, 0.9903, 0.9959, 1.0000, 1.0031, 1.0055, 1.0072,
        1.0085, 1.0095, 1.0103, 1.0108, 1.0112, 1.0116, 1.1018])
finalReturns:  tensor([7.0088, 7.0988, 7.0459, 6.8892, 6.6567, 6.3684, 6.0389,
5.6789, 5.2960,
        4.8963, 4.4839, 4.0621, 3.6332, 3.1990, 2.7608, 2.3197, 1.8764, 1.4313,
        0.9851, 0.5379])
----------------------------------------
iter  0  stage  4  ep  0   adversary:  AdversaryModes.constant_132
  actions:  tensor([30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30,
30, 30, 30,
        30, 30, 30, 30, 30, 30,  0])
```

```
loss=  tensor(0.0616, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([13.0598, 13.0598, 13.0598, 13.0598, 13.0598, 12.1686, 11.3667, 10.6286,
         9.9366,  9.2783,  8.6446,  8.0291,  7.4271,  6.8351,  6.2506,  5.6715,
         5.0967,  4.5249,  3.9555,  3.3878,  2.8214,  2.2560,  1.6913,  1.1272,
         0.5634]) return=  228471.31797735966
probs of actions:  tensor([0.9991, 0.9989, 0.9985, 0.9994, 0.9992, 0.9990,
0.9994, 0.9994, 0.9995,
        0.9998, 0.9995, 0.9996, 0.9993, 0.9990, 0.9997, 0.9993, 0.9995, 0.9990,
        0.9998, 1.0000, 1.0000, 1.0000, 0.9999, 0.9921, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4212, 0.5480, 0.6523, 0.7357, 0.8011, 0.8519, 0.8908,
0.9206, 0.9432,
        0.9603, 0.9732, 0.9830, 0.9903, 0.9959, 1.0000, 1.0031, 1.0055, 1.0072,
        1.0085, 1.0095, 1.0103, 1.0108, 1.0112, 1.0116, 1.1018])
finalReturns:  tensor([7.4301, 7.5201, 7.4702, 7.3174, 7.0888, 6.8040, 6.4773,
6.1196, 5.7386,
        5.3403, 4.9290, 4.5080, 4.0798, 3.6460, 3.2083, 2.7674, 2.3243, 1.8794,
        1.4333, 0.9862, 0.5384])
----------------------------------------
iter  0  stage  3  ep  0   adversary:  AdversaryModes.constant_132
  actions:  tensor([30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30,
30, 30, 30,
        30, 30, 30, 30, 30, 30,  0])
loss=  tensor(0.0675, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([13.3829, 13.3829, 13.3829, 13.3829, 12.5572, 11.8020, 11.0977, 10.4303,
         9.7900,  9.1696,  8.5639,  7.9692,  7.3826,  6.8021,  6.2261,  5.6535,
         5.0835,  4.5153,  3.9485,  3.3828,  2.8180,  2.2537,  1.6898,  1.1263,
         0.5631]) return=  228471.31797735966
probs of actions:  tensor([0.9991, 0.9989, 0.9985, 0.9994, 0.9992, 0.9990,
0.9994, 0.9994, 0.9995,
        0.9998, 0.9995, 0.9996, 0.9993, 0.9990, 0.9997, 0.9993, 0.9995, 0.9990,
        0.9998, 1.0000, 1.0000, 1.0000, 0.9999, 0.9921, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4212, 0.5480, 0.6523, 0.7357, 0.8011, 0.8519, 0.8908,
0.9206, 0.9432,
        0.9603, 0.9732, 0.9830, 0.9903, 0.9959, 1.0000, 1.0031, 1.0055, 1.0072,
        1.0085, 1.0095, 1.0103, 1.0108, 1.0112, 1.0116, 1.1018])
finalReturns:  tensor([7.8427, 7.9327, 7.8868, 7.7392, 7.5158, 7.2355, 6.9127,
6.5581, 6.1795,
        5.7831, 5.3733, 4.9534, 4.5260, 4.0930, 3.6557, 3.2152, 2.7724, 2.3277,
        1.8818, 1.4348, 0.9870, 0.5387])
----------------------------------------
iter  0  stage  2  ep  258   adversary:  AdversaryModes.constant_132
  actions:  tensor([30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30,
26, 30, 30,
        30, 30, 30, 30, 30, 30,  0])
loss=  tensor(34.8480, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([13.6344, 13.6344, 13.6344, 12.8921, 12.1971, 11.5366, 10.9012, 10.2846,
```

```
         9.6816,  9.0890,  8.5039,  7.9245,  7.3494,  6.7774,  6.2078,  5.6400,
         5.0735,  4.5081,  3.9433,  3.3791,  2.8154,  2.2519,  1.6887,  1.1257,
         0.5628]) return=  227922.18817761683
probs of actions:  tensor([9.9912e-01, 9.9896e-01, 9.9900e-01, 9.9960e-01,
9.9942e-01, 9.9922e-01,
        9.9947e-01, 9.9954e-01, 9.9955e-01, 9.9982e-01, 9.9957e-01, 9.9970e-01,
        9.9947e-01, 9.9904e-01, 9.9969e-01, 4.2794e-04, 9.9954e-01, 9.9905e-01,
        9.9980e-01, 9.9997e-01, 1.0000e+00, 1.0000e+00, 9.9995e-01, 9.9227e-01,
        1.0000e+00], grad_fn=<ExpBackward0>)
rewards:  tensor([0.4212, 0.5480, 0.6523, 0.7357, 0.8011, 0.8519, 0.8908,
0.9206, 0.9432,
        0.9603, 0.9732, 0.9830, 0.9903, 0.9959, 1.0000, 1.0255, 0.9846, 0.9916,
        0.9968, 1.0007, 1.0036, 1.0059, 1.0075, 1.0088, 1.0997])
finalReturns:  tensor([8.1886, 8.2786, 8.2379, 8.0973, 7.8807, 7.6066, 7.2889,
6.9384, 6.5631,
        6.1693, 5.7614, 5.3431, 4.9168, 4.4846, 4.0256, 3.6064, 3.1796, 2.7470,
        2.3101, 1.8699, 1.4272, 0.9828, 0.5369])
----------------------------------------
iter  0  stage  1  ep  23   adversary:  AdversaryModes.constant_132
  actions:  tensor([30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30,
30, 30, 30,
        30, 30, 30, 30, 30, 30,  0])
loss=  tensor(0.0766, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([13.7976, 13.7976, 13.1596, 12.5410, 11.9366, 11.3428, 10.7569, 10.1769,
         9.6013,  9.0290,  8.4592,  7.8912,  7.3246,  6.7590,  6.1942,  5.6299,
         5.0661,  4.5026,  3.9394,  3.3764,  2.8134,  2.2506,  1.6879,  1.1252,
         0.5626]) return=  228471.31797735966
probs of actions:  tensor([0.9991, 0.9990, 0.9990, 0.9996, 0.9994, 0.9992,
0.9995, 0.9995, 0.9996,
        0.9998, 0.9996, 0.9997, 0.9995, 0.9990, 0.9997, 0.9993, 0.9995, 0.9991,
        0.9998, 1.0000, 1.0000, 1.0000, 0.9999, 0.9922, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4212, 0.5480, 0.6523, 0.7357, 0.8011, 0.8519, 0.8908,
0.9206, 0.9432,
        0.9603, 0.9732, 0.9830, 0.9903, 0.9959, 1.0000, 1.0031, 1.0055, 1.0072,
        1.0085, 1.0095, 1.0103, 1.0108, 1.0112, 1.0116, 1.1018])
finalReturns:  tensor([8.6283, 8.7183, 8.6847, 8.5534, 8.3460, 8.0800, 7.7691,
7.4241, 7.0532,
        6.6628, 6.2575, 5.8412, 5.4164, 4.9854, 4.5496, 4.1103, 3.6683, 3.2244,
        2.7789, 2.3323, 1.8848, 1.4367, 0.9881, 0.5392])
----------------------------------------
iter  0  stage  0  ep  0   adversary:  AdversaryModes.constant_132
  actions:  tensor([30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30,
30, 30, 30,
        30, 30, 30, 30, 30, 30,  0])
loss=  tensor(0.0858, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([13.8555, 13.3442, 12.8204, 12.2871, 11.7465, 11.2005, 10.6504, 10.0972,
         9.5417,  8.9844,  8.4258,  7.8663,  7.3060,  6.7452,  6.1839,  5.6224,
```

```
         5.0606,  4.4986,  3.9365,  3.3743,  2.8120,  2.2497,  1.6873,  1.1249,
         0.5624]) return=  228471.31797735966
probs of actions:  tensor([0.9991, 0.9990, 0.9990, 0.9996, 0.9994, 0.9992,
0.9995, 0.9995, 0.9996,
        0.9998, 0.9996, 0.9997, 0.9995, 0.9990, 0.9997, 0.9993, 0.9995, 0.9991,
        0.9998, 1.0000, 1.0000, 1.0000, 0.9999, 0.9922, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4212, 0.5480, 0.6523, 0.7357, 0.8011, 0.8519, 0.8908,
0.9206, 0.9432,
        0.9603, 0.9732, 0.9830, 0.9903, 0.9959, 1.0000, 1.0031, 1.0055, 1.0072,
        1.0085, 1.0095, 1.0103, 1.0108, 1.0112, 1.0116, 1.1018])
finalReturns:  tensor([8.9917, 9.0817, 9.0575, 8.9385, 8.7434, 8.4883, 8.1865,
7.8489, 7.4838,
        7.0979, 6.6961, 6.2824, 5.8597, 5.4302, 4.9956, 4.5572, 4.1159, 3.6724,
        3.2273, 2.7810, 2.3337, 1.8858, 1.4373, 0.9885, 0.5393])
0,[1e-05,1][1, 10000, 1, 1],1682476508 saved
[830550, 'tensor([0., 1., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.])',
228471.31797735966, 18758.68173295266, 0.08583260327577591, 1e-05, 1, 0,
'tensor([30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30,
30,\n        30, 30, 30, 30, 30, 30,  0])', '[1.   1.   1.   1.   1.   1.   1.
1.   1.   1.   1.   1.   1.   1.   1.\n 1.   1.   1.   1.   1.   1.   1.   1.
0.99 1.  ]', '0,[1e-05,1][1, 10000, 1, 1],1682476508', 25, 50,
174696.33353181678, 228471.31797735966, 95930.56714139864, 130490.57866666671,
127475.47466666666, 61768.648600928995, 61760.04937694948, 76910.22725950743,
76910.22725950743, 110767.6026843345, 61768.648600928995, 76910.22725950743]
policy reset
----------------------------------------
iter  1  stage  24  ep  99999   adversary:  AdversaryModes.constant_132
  actions:  tensor([ 0,  0,  0,  0,  0,  3,  0,  0,  0,  0,  0,  0, 14,  0, 23,
0,  0,  0,
         0,  0,  0,  0,  0,  0,  0])
loss=  tensor(-0., grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.5712,
0.5712, 0.5712, 0.5712, 0.5712, 0.5712, 0.5712, 0.5712, 0.5712,
        0.5712, 0.5712, 0.5712, 0.5712, 0.5712, 0.5712, 0.5712, 0.5712, 0.5712,
        0.5712, 0.5712, 0.5712, 0.5712, 0.5712, 0.5712, 0.5712]) return=
143712.4363256526
probs of actions:  tensor([9.1195e-01, 9.0448e-01, 9.2248e-01, 9.3650e-01,
9.1132e-01, 4.6668e-03,
        9.2003e-01, 9.1205e-01, 8.9672e-01, 9.0821e-01, 9.1395e-01, 9.0437e-01,
        4.1970e-04, 9.3013e-01, 4.8654e-04, 9.3345e-01, 9.3245e-01, 9.2076e-01,
        9.3494e-01, 9.0264e-01, 9.2735e-01, 9.1623e-01, 9.3351e-01, 9.3845e-01,
        9.9369e-01], grad_fn=<ExpBackward0>)
rewards:  tensor([0.5112, 0.5238, 0.5334, 0.5406, 0.5460, 0.5492, 0.5644,
0.5639, 0.5636,
        0.5633, 0.5631, 0.5630, 0.5432, 0.6165, 0.5499, 0.6844, 0.6528, 0.6296,
        0.6125, 0.5998, 0.5903, 0.5833, 0.5781, 0.5742, 0.5712])
finalReturns:  tensor([0.])
----------------------------------------
```

```
iter  1  stage  23  ep  99999   adversary:  AdversaryModes.constant_132
  actions:  tensor([ 2, 13,  8,  3,  8, 11,  1,  0, 10,  1,  4,  0, 11, 10, 11,
11,  0, 11,
         3, 19,  5, 11,  0, 16,  0])
loss=  tensor(0.1585, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.3100,
1.3100, 1.3100, 1.3100, 1.3100, 1.3100, 1.3100, 1.3100, 1.3100,
        1.3100, 1.3100, 1.3100, 1.3100, 1.3100, 1.3100, 1.3100, 1.3100, 1.3100,
        1.3100, 1.3100, 1.3100, 1.3100, 1.3100, 1.3100, 0.6413]) return=
158781.70950982286
probs of actions:  tensor([0.1258, 0.0380, 0.0482, 0.0496, 0.0497, 0.0659,
0.0877, 0.3758, 0.0628,
        0.0816, 0.0299, 0.1993, 0.0989, 0.0494, 0.1139, 0.0767, 0.3202, 0.0930,
        0.0514, 0.0096, 0.0470, 0.1015, 0.3275, 0.0204, 0.9615],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5108, 0.5142, 0.5812, 0.6113, 0.6048, 0.6181, 0.6566,
0.6365, 0.6076,
        0.6429, 0.6247, 0.6258, 0.5975, 0.6310, 0.6488, 0.6680, 0.6948, 0.6483,
        0.6789, 0.6255, 0.7116, 0.6831, 0.7061, 0.6431, 0.7070])
finalReturns:  tensor([0.0401, 0.0657])
----------------------------------------
iter  1  stage  22  ep  99999   adversary:  AdversaryModes.constant_132
  actions:  tensor([17, 17, 22,  9, 27, 10, 23,  1, 11, 22, 19, 22, 11, 22, 23,
17, 21, 19,
        14, 19,  1, 22, 27, 21,  0])
loss=  tensor(0.9555, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.2835,
2.2835, 2.2835, 2.2835, 2.2835, 2.2835, 2.2835, 2.2835, 2.2835,
        2.2835, 2.2835, 2.2835, 2.2835, 2.2835, 2.2835, 2.2835, 2.2835, 2.2835,
        2.2835, 2.2835, 2.2835, 2.2835, 2.2835, 1.4586, 0.7040]) return=
189588.30780201865
probs of actions:  tensor([0.0446, 0.0415, 0.1311, 0.0181, 0.0726, 0.0367,
0.0968, 0.0634, 0.0767,
        0.1486, 0.0753, 0.1825, 0.1011, 0.2034, 0.0902, 0.0444, 0.0571, 0.0620,
        0.0055, 0.0769, 0.0473, 0.1604, 0.1186, 0.1233, 0.9659],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4823, 0.5582, 0.5991, 0.7076, 0.6403, 0.7792, 0.7194,
0.8173, 0.7414,
        0.7016, 0.7596, 0.7689, 0.8353, 0.7713, 0.7963, 0.8475, 0.8247, 0.8458,
        0.8627, 0.8232, 0.8652, 0.7396, 0.7520, 0.8324, 0.8876])
finalReturns:  tensor([0.1885, 0.2614, 0.1837])
----------------------------------------
iter  1  stage  21  ep  99999   adversary:  AdversaryModes.constant_132
  actions:  tensor([ 2, 23, 29, 27, 27, 33, 22, 22, 29, 27, 27, 27, 29, 27, 27,
23, 19, 23,
        27, 27, 29, 29, 21, 29,  0])
loss=  tensor(2.9230, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.4631,
3.4631, 3.4631, 3.4631, 3.4631, 3.4631, 3.4631, 3.4631, 3.4631,
        3.4631, 3.4631, 3.4631, 3.4631, 3.4631, 3.4631, 3.4631, 3.4631, 3.4631,
        3.4631, 3.4631, 3.4631, 3.4631, 2.4400, 1.5449, 0.7401]) return=
```

```
213779.62568054372
probs of actions:  tensor([0.0017, 0.1390, 0.2575, 0.2755, 0.2901, 0.0024,
0.1132, 0.1030, 0.2525,
        0.3007, 0.2977, 0.2952, 0.2668, 0.3061, 0.2577, 0.1412, 0.0256, 0.1252,
        0.2708, 0.2615, 0.2655, 0.2885, 0.0294, 0.2154, 0.9868],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5108, 0.4782, 0.5425, 0.6559, 0.7286, 0.7494, 0.8825,
0.8850, 0.8512,
        0.8980, 0.9151, 0.9281, 0.9267, 0.9554, 0.9584, 0.9807, 0.9789, 0.9284,
        0.9032, 0.9191, 0.9199, 0.9390, 0.9935, 0.9238, 1.0261])
finalReturns:  tensor([0.4193, 0.5034, 0.4050, 0.2860])
---------------------------------------
iter  1  stage  20  ep  99999   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
27, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.3539, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.3058,
4.3058, 4.3058, 4.3058, 4.3058, 4.3058, 4.3058, 4.3058, 4.3058, 4.3058,
        4.3058, 4.3058, 4.3058, 4.3058, 4.3058, 4.3058, 4.3058, 4.3058, 4.3058,
        4.3058, 4.3058, 4.3058, 3.2296, 2.2973, 1.4662, 0.7072]) return=
225339.17689685564
probs of actions:  tensor([0.8154, 0.8062, 0.8327, 0.8258, 0.8574, 0.8441,
0.8384, 0.8128, 0.8276,
        0.8399, 0.8353, 0.8444, 0.8512, 0.8138, 0.8433, 0.0926, 0.8387, 0.8597,
        0.8749, 0.8681, 0.9106, 0.8920, 0.9257, 0.8106, 0.9973],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9997, 0.9804, 0.9847,
        0.9879, 0.9903, 0.9921, 0.9934, 0.9944, 0.9952, 1.0799])
finalReturns:  tensor([0.7493, 0.8334, 0.7722, 0.6089, 0.3727])
---------------------------------------
iter  1  stage  19  ep  99999   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 27, 29, 29, 29, 29, 29, 29, 29, 29, 27,
29, 29, 29,
        29, 27, 29, 29, 29, 29,  0])
loss=  tensor(3.4990, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.9724,
4.9724, 4.9724, 4.9724, 4.9724, 4.9724, 4.9724, 4.9724, 4.9724, 4.9724,
        4.9724, 4.9724, 4.9724, 4.9724, 4.9724, 4.9724, 4.9724, 4.9724, 4.9724,
        4.9724, 4.9724, 3.8972, 2.9656, 2.1349, 1.3762, 0.6693]) return=
224839.01113301513
probs of actions:  tensor([0.9082, 0.9036, 0.9186, 0.9093, 0.9285, 0.0573,
0.9221, 0.9087, 0.9121,
        0.9198, 0.9195, 0.9243, 0.9271, 0.9111, 0.0542, 0.9275, 0.9204, 0.9342,
        0.9394, 0.0433, 0.9618, 0.9606, 0.9649, 0.9095, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8542, 0.8707,
0.9017, 0.9252,
```

```
        0.9431, 0.9565, 0.9667, 0.9744, 0.9801, 0.9957, 0.9774, 0.9824, 0.9862,
        0.9890, 1.0023, 0.9824, 0.9861, 0.9890, 0.9911, 1.0768])
finalReturns:  tensor([1.0553, 1.1282, 1.0774, 0.9220, 0.6917, 0.4075])
----------------------------------------
iter  1  stage  18  ep  99999   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.1558, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.6221,
5.6221, 5.6221, 5.6221, 5.6221, 5.6221, 5.6221, 5.6221, 5.6221,
        5.6221, 5.6221, 5.6221, 5.6221, 5.6221, 5.6221, 5.6221, 5.6221, 5.6221,
        5.6221, 4.5443, 3.6109, 2.7790, 2.0194, 1.3118, 0.6421]) return=
225610.63312213228
probs of actions:  tensor([0.9589, 0.9574, 0.9638, 0.9603, 0.9687, 0.9671,
0.9659, 0.9609, 0.9606,
        0.9646, 0.9645, 0.9679, 0.9693, 0.9613, 0.9680, 0.9688, 0.9643, 0.9729,
        0.9797, 0.9749, 0.9883, 0.9855, 0.9864, 0.9523, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([1.4313, 1.5154, 1.4541, 1.2907, 1.0544, 0.7657, 0.4389])
----------------------------------------
iter  1  stage  17  ep  99999   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0674, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.2397,
6.2397, 6.2397, 6.2397, 6.2397, 6.2397, 6.2397, 6.2397, 6.2397,
        6.2397, 6.2397, 6.2397, 6.2397, 6.2397, 6.2397, 6.2397, 6.2397, 6.2397,
        5.1632, 4.2307, 3.3994, 2.6402, 1.9330, 1.2634, 0.6215]) return=
225610.63312213228
probs of actions:  tensor([0.9847, 0.9841, 0.9869, 0.9858, 0.9886, 0.9888,
0.9882, 0.9863, 0.9859,
        0.9873, 0.9876, 0.9891, 0.9897, 0.9864, 0.9889, 0.9893, 0.9873, 0.9919,
        0.9960, 0.9936, 0.9962, 0.9956, 0.9956, 0.9809, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([1.8061, 1.8902, 1.8290, 1.6657, 1.4295, 1.1409, 0.8141,
0.4594])
----------------------------------------
iter  1  stage  16  ep  85058   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
```

```
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0069, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.8407,
6.8407, 6.8407, 6.8407, 6.8407, 6.8407, 6.8407, 6.8407, 6.8407,
        6.8407, 6.8407, 6.8407, 6.8407, 6.8407, 6.8407, 6.8407, 6.8407, 5.7658,
        4.8345, 4.0040, 3.2454, 2.5386, 1.8694, 1.2278, 0.6064]) return=
225610.63312213228
probs of actions:  tensor([0.9980, 0.9978, 0.9983, 0.9983, 0.9986, 0.9987,
0.9986, 0.9983, 0.9983,
        0.9985, 0.9986, 0.9988, 0.9989, 0.9985, 0.9988, 0.9988, 0.9990, 0.9992,
        0.9998, 0.9997, 1.0000, 0.9997, 0.9997, 0.9987, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([2.1959, 2.2800, 2.2189, 2.0557, 1.8196, 1.5311, 1.2044,
0.8498, 0.4745])
---------------------------------------
iter  1  stage  15  ep  2201   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0087, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.4283,
7.4283, 7.4283, 7.4283, 7.4283, 7.4283, 7.4283, 7.4283, 7.4283,
        7.4283, 7.4283, 7.4283, 7.4283, 7.4283, 7.4283, 7.4283, 6.3558, 5.4260,
        4.5967, 3.8389, 3.1326, 2.4638, 1.8225, 1.2013, 0.5951]) return=
225610.63312213228
probs of actions:  tensor([0.9983, 0.9982, 0.9985, 0.9985, 0.9988, 0.9989,
0.9988, 0.9985, 0.9985,
        0.9987, 0.9988, 0.9990, 0.9991, 0.9987, 0.9990, 0.9990, 0.9993, 0.9993,
        0.9999, 0.9998, 1.0000, 0.9998, 0.9998, 0.9989, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([2.5967, 2.6808, 2.6198, 2.4567, 2.2208, 1.9324, 1.6059,
1.2513, 0.8762,
        0.4858])
---------------------------------------
iter  1  stage  14  ep  15   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0127, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.0050,
8.0050, 8.0050, 8.0050, 8.0050, 8.0050, 8.0050, 8.0050, 8.0050,
        8.0050, 8.0050, 8.0050, 8.0050, 8.0050, 8.0050, 6.9354, 6.0078, 5.1799,
        4.4232, 3.7177, 3.0495, 2.4086, 1.7877, 1.1817, 0.5868]) return=
```

```
225610.63312213228
probs of actions:  tensor([0.9983, 0.9982, 0.9986, 0.9986, 0.9988, 0.9989,
0.9988, 0.9985, 0.9985,
        0.9987, 0.9988, 0.9990, 0.9991, 0.9987, 0.9990, 0.9990, 0.9993, 0.9993,
        0.9999, 0.9998, 1.0000, 0.9998, 0.9998, 0.9989, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([3.0055, 3.0896, 3.0288, 2.8659, 2.6302, 2.3420, 2.0156,
1.6611, 1.2861,
        0.8958, 0.4942])
----------------------------------------
iter  1  stage  13  ep  280   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0172, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.5720,
8.5720, 8.5720, 8.5720, 8.5720, 8.5720, 8.5720, 8.5720, 8.5720,
        8.5720, 8.5720, 8.5720, 8.5720, 8.5720, 7.5064, 6.5815, 5.7556, 5.0003,
        4.2958, 3.6284, 2.9880, 2.3676, 1.7619, 1.1671, 0.5805]) return=
225610.63312213228
probs of actions:  tensor([0.9983, 0.9981, 0.9985, 0.9985, 0.9988, 0.9989,
0.9988, 0.9985, 0.9985,
        0.9987, 0.9988, 0.9990, 0.9991, 0.9990, 0.9991, 0.9990, 0.9993, 0.9993,
        0.9999, 0.9998, 1.0000, 0.9998, 0.9998, 0.9989, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([3.4201, 3.5042, 3.4436, 3.2810, 3.0455, 2.7576, 2.4313,
2.0771, 1.7022,
        1.3119, 0.9104, 0.5004])
----------------------------------------
iter  1  stage  12  ep  0   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0222, grad_fn=<NegBackward0>)   ,  base rewards= tensor([9.1296,
9.1296, 9.1296, 9.1296, 9.1296, 9.1296, 9.1296, 9.1296, 9.1296,
        9.1296, 9.1296, 9.1296, 9.1296, 8.0693, 7.1481, 6.3249, 5.5714, 4.8684,
        4.2019, 3.5622, 2.9423, 2.3370, 1.7426, 1.1562, 0.5759]) return=
225610.63312213228
probs of actions:  tensor([0.9983, 0.9981, 0.9985, 0.9985, 0.9988, 0.9989,
0.9988, 0.9985, 0.9985,
        0.9987, 0.9988, 0.9990, 0.9991, 0.9990, 0.9991, 0.9990, 0.9993, 0.9993,
```

```
        0.9999, 0.9998, 1.0000, 0.9998, 0.9998, 0.9989, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([3.8386, 3.9227, 3.8624, 3.7002, 3.4651, 3.1775, 2.8515,
2.4975, 2.1227,
        1.7327, 1.3312, 0.9213, 0.5051])
---------------------------------------
iter  1  stage  11  ep  2   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0279, grad_fn=<NegBackward0>)   ,  base rewards= tensor([9.6775,
9.6775, 9.6775, 9.6775, 9.6775, 9.6775, 9.6775, 9.6775, 9.6775,
        9.6775, 9.6775, 9.6775, 8.6243, 7.7080, 6.8883, 6.1374, 5.4361, 4.7710,
        4.1323, 3.5131, 2.9084, 2.3143, 1.7282, 1.1481, 0.5724]) return=
225610.63312213228
probs of actions:  tensor([0.9983, 0.9981, 0.9985, 0.9985, 0.9988, 0.9989,
0.9988, 0.9985, 0.9985,
        0.9987, 0.9988, 0.9990, 0.9991, 0.9990, 0.9991, 0.9990, 0.9993, 0.9993,
        0.9999, 0.9998, 1.0000, 0.9998, 0.9998, 0.9989, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([4.2598, 4.3439, 4.2840, 4.1223, 3.8877, 3.6005, 3.2749,
2.9211, 2.5466,
        2.1567, 1.7554, 1.3456, 0.9294, 0.5085])
---------------------------------------
iter  1  stage  10  ep  14016   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0265, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([10.2146, 10.2146, 10.2146, 10.2146, 10.2146, 10.2146, 10.2146, 10.2146,
        10.2146, 10.2146, 10.2146,  9.1707,  8.2611,  7.4460,  6.6984,  5.9995,
         5.3362,  4.6988,  4.0806,  3.4765,  2.8830,  2.2973,  1.7175,  1.1420,
         0.5698]) return=  225610.63312213228
probs of actions:  tensor([0.9982, 0.9981, 0.9985, 0.9985, 0.9987, 0.9989,
0.9987, 0.9984, 0.9984,
        0.9987, 0.9990, 1.0000, 0.9997, 0.9990, 0.9990, 0.9989, 0.9993, 0.9993,
        1.0000, 0.9998, 1.0000, 0.9998, 0.9998, 0.9989, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
```

```
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([4.6825, 4.7666, 4.7072, 4.5461, 4.3122, 4.0255, 3.7004,
3.3470, 2.9728,
        2.5832, 2.1820, 1.7724, 1.3563, 0.9355, 0.5111])
----------------------------------------
iter  1  stage  9  ep  432   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0336, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([10.7388, 10.7388, 10.7388, 10.7388, 10.7388, 10.7388, 10.7388, 10.7388,
        10.7388, 10.7388,  9.7074,  8.8064,  7.9975,  7.2544,  6.5588,  5.8977,
         5.2621,  4.6452,  4.0421,  3.4493,  2.8641,  2.2847,  1.7095,  1.1375,
         0.5678]) return=  225610.63312213228
probs of actions:  tensor([0.9981, 0.9980, 0.9984, 0.9984, 0.9987, 0.9988,
0.9987, 0.9984, 0.9984,
        0.9990, 0.9989, 1.0000, 0.9997, 0.9990, 0.9990, 0.9988, 0.9993, 0.9993,
        1.0000, 0.9998, 1.0000, 0.9997, 0.9998, 0.9988, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([5.1056, 5.1897, 5.1309, 4.9707, 4.7377, 4.4518, 4.1273,
3.7745, 3.4007,
        3.0113, 2.6104, 2.2010, 1.7851, 1.3643, 0.9401, 0.5131])
----------------------------------------
iter  1  stage  8  ep  6551   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0253, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([11.2472, 11.2472, 11.2472, 11.2472, 11.2472, 11.2472, 11.2472, 11.2472,
        11.2472, 10.2322,  9.3428,  8.5420,  7.8048,  7.1135,  6.4556,  5.8223,
         5.2071,  4.6052,  4.0134,  3.4289,  2.8500,  2.2752,  1.7035,  1.1341,
         0.5664]) return=  225610.63312213228
probs of actions:  tensor([0.9987, 0.9987, 0.9989, 0.9990, 0.9991, 0.9992,
0.9991, 0.9989, 0.9990,
        0.9995, 0.9994, 1.0000, 0.9998, 0.9994, 0.9994, 0.9993, 0.9997, 0.9996,
        1.0000, 0.9999, 1.0000, 0.9999, 0.9999, 0.9993, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([5.5282, 5.6123, 5.5544, 5.3953, 5.1634, 4.8785, 4.5549,
4.2028, 3.8295,
```

```
        3.4406, 3.0400, 2.6308, 2.2151, 1.7945, 1.3704, 0.9435, 0.5145])
---------------------------------------
iter  1  stage  7  ep  30   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0320, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([11.7354, 11.7354, 11.7354, 11.7354, 11.7354, 11.7354, 11.7354, 11.7354,
        10.7421,  9.8680,  9.0781,  8.3488,  7.6631,  7.0094,  6.3791,  5.7662,
         5.1660,  4.5754,  3.9919,  3.4137,  2.8394,  2.2681,  1.6990,  1.1315,
         0.5653]) return=  225610.63312213228
probs of actions:  tensor([0.9988, 0.9987, 0.9989, 0.9990, 0.9992, 0.9992,
0.9991, 0.9990, 0.9990,
        0.9995, 0.9994, 1.0000, 0.9998, 0.9994, 0.9994, 0.9993, 0.9997, 0.9996,
        1.0000, 0.9999, 1.0000, 0.9999, 0.9999, 0.9993, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([5.9491, 6.0332, 5.9765, 5.8190, 5.5886, 5.3051, 4.9826,
4.6314, 4.2589,
        3.8705, 3.4704, 3.0615, 2.6460, 2.2257, 1.8016, 1.3749, 0.9460, 0.5156])
---------------------------------------
iter  1  stage  6  ep  0   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0387, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([12.1977, 12.1977, 12.1977, 12.1977, 12.1977, 12.1977, 12.1977, 11.2331,
        10.3791,  9.6037,  8.8847,  8.2066,  7.5584,  6.9322,  6.3223,  5.7244,
         5.1354,  4.5532,  3.9759,  3.4023,  2.8315,  2.2627,  1.6956,  1.1296,
         0.5645]) return=  225610.63312213228
probs of actions:  tensor([0.9988, 0.9987, 0.9989, 0.9990, 0.9992, 0.9992,
0.9991, 0.9990, 0.9990,
        0.9995, 0.9994, 1.0000, 0.9998, 0.9994, 0.9994, 0.9993, 0.9997, 0.9996,
        1.0000, 0.9999, 1.0000, 0.9999, 0.9999, 0.9993, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([6.3673, 6.4514, 6.3962, 6.2408, 6.0125, 5.7308, 5.4098,
5.0598, 4.6882,
        4.3007, 3.9011, 3.4927, 3.0775, 2.6574, 2.2336, 1.8070, 1.3782, 0.9479,
        0.5165])
---------------------------------------
iter  1  stage  5  ep  0   adversary:  AdversaryModes.constant_132
```

```
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0451, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([12.6268, 12.6268, 12.6268, 12.6268, 12.6268, 12.6268, 11.6997, 10.8722,
        10.1157,  9.4104,  8.7424,  8.1016,  7.4808,  6.8749,  6.2799,  5.6932,
         5.1126,  4.5365,  3.9639,  3.3937,  2.8255,  2.2588,  1.6931,  1.1282,
         0.5639]) return=  225610.63312213228
probs of actions:  tensor([0.9988, 0.9987, 0.9990, 0.9990, 0.9992, 0.9992,
0.9991, 0.9990, 0.9990,
        0.9995, 0.9994, 1.0000, 0.9998, 0.9994, 0.9994, 0.9993, 0.9997, 0.9996,
        1.0000, 0.9999, 1.0000, 0.9999, 0.9999, 0.9993, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([6.7813, 6.8654, 6.8123, 6.6596, 6.4340, 6.1547, 5.8358,
5.4874, 5.1172,
        4.7306, 4.3318, 3.9240, 3.5093, 3.0895, 2.6660, 2.2395, 1.8109, 1.3808,
        0.9494, 0.5171])
----------------------------------------
iter  1  stage  4  ep  0   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0525, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([13.0129, 13.0129, 13.0129, 13.0129, 13.0129, 12.1346, 11.3419, 10.6103,
         9.9232,  9.2683,  8.6372,  8.0237,  7.4231,  6.8321,  6.2483,  5.6699,
         5.0955,  4.5241,  3.9549,  3.3874,  2.8211,  2.2558,  1.6912,  1.1271,
         0.5634]) return=  225610.63312213228
probs of actions:  tensor([0.9988, 0.9987, 0.9990, 0.9990, 0.9992, 0.9992,
0.9991, 0.9990, 0.9990,
        0.9995, 0.9994, 1.0000, 0.9998, 0.9994, 0.9994, 0.9993, 0.9997, 0.9996,
        1.0000, 0.9999, 1.0000, 0.9999, 0.9999, 0.9993, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([7.1893, 7.2734, 7.2232, 7.0741, 6.8522, 6.5761, 6.2599,
5.9137, 5.5452,
        5.1599, 4.7622, 4.3551, 3.9411, 3.5218, 3.0985, 2.6724, 2.2440, 1.8139,
        1.3826, 0.9504, 0.5175])
----------------------------------------
iter  1  stage  3  ep  14   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
```

```
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0604, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([13.3441, 13.3441, 13.3441, 13.3441, 12.5289, 11.7813, 11.0824, 10.4190,
         9.7816,  9.1633,  8.5593,  7.9657,  7.3800,  6.8002,  6.2247,  5.6525,
         5.0827,  4.5147,  3.9481,  3.3826,  2.8178,  2.2535,  1.6897,  1.1263,
         0.5630]) return=  225610.63312213228
probs of actions:  tensor([0.9988, 0.9987, 0.9990, 0.9990, 0.9992, 0.9993,
0.9992, 0.9990, 0.9990,
        0.9995, 0.9995, 1.0000, 0.9998, 0.9994, 0.9994, 0.9993, 0.9997, 0.9996,
        1.0000, 0.9999, 1.0000, 0.9999, 0.9999, 0.9993, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([7.5892, 7.6733, 7.6268, 7.4827, 7.2655, 6.9938, 6.6811,
6.3379, 5.9716,
        5.5882, 5.1918, 4.7858, 4.3726, 3.9539, 3.5311, 3.1053, 2.6771, 2.2473,
        1.8162, 1.3841, 0.9512, 0.5179])
----------------------------------------
iter  1  stage  2  ep  19   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0680, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([13.6060, 13.6060, 13.6060, 12.8713, 12.1817, 11.5251, 10.8928, 10.2782,
         9.6769,  9.0855,  8.5013,  7.9226,  7.3479,  6.7763,  6.2070,  5.6394,
         5.0731,  4.5077,  3.9431,  3.3790,  2.8153,  2.2519,  1.6887,  1.1257,
         0.5628]) return=  225610.63312213228
probs of actions:  tensor([0.9988, 0.9987, 0.9990, 0.9990, 0.9992, 0.9993,
0.9992, 0.9991, 0.9990,
        0.9995, 0.9995, 1.0000, 0.9998, 0.9994, 0.9994, 0.9993, 0.9997, 0.9996,
        1.0000, 0.9999, 1.0000, 0.9999, 0.9999, 0.9993, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([7.9780, 8.0621, 8.0205, 7.8829, 7.6723, 7.4063, 7.0984,
6.7590, 6.3958,
        6.0148, 5.6203, 5.2157, 4.8035, 4.3856, 3.9635, 3.5381, 3.1103, 2.6807,
        2.2498, 1.8178, 1.3851, 0.9518, 0.5181])
----------------------------------------
iter  1  stage  1  ep  2390   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0622, grad_fn=<NegBackward0>)   ,  base rewards=
```

```
tensor([13.7821, 13.7821, 13.1481, 12.5324, 11.9302, 11.3380, 10.7533, 10.1743,
         9.5993,  9.0276,  8.4581,  7.8904,  7.3239,  6.7585,  6.1938,  5.6297,
         5.0659,  4.5025,  3.9393,  3.3763,  2.8134,  2.2506,  1.6879,  1.1252,
         0.5626]) return=  225610.63312213228
probs of actions:  tensor([0.9989, 0.9990, 0.9991, 0.9992, 0.9994, 0.9997,
0.9997, 0.9993, 0.9991,
        0.9995, 0.9995, 1.0000, 0.9999, 0.9995, 0.9995, 0.9994, 0.9998, 0.9996,
        1.0000, 0.9999, 1.0000, 0.9999, 0.9999, 0.9994, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([8.3518, 8.4359, 8.4010, 8.2721, 8.0700, 7.8117, 7.5102,
7.1760, 6.8169,
        6.4390, 6.0470, 5.6443, 5.2335, 4.8167, 4.3954, 3.9706, 3.5433, 3.1141,
        2.6834, 2.2517, 1.8191, 1.3859, 0.9523, 0.5183])
----------------------------------------
iter  1  stage  0  ep  100   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0722, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([13.8555, 13.3442, 12.8204, 12.2871, 11.7465, 11.2005, 10.6504, 10.0972,
         9.5417,  8.9844,  8.4258,  7.8663,  7.3060,  6.7452,  6.1839,  5.6224,
         5.0606,  4.4986,  3.9365,  3.3743,  2.8120,  2.2497,  1.6873,  1.1249,
         0.5624]) return=  225610.63312213228
probs of actions:  tensor([0.9990, 0.9990, 0.9991, 0.9992, 0.9994, 0.9997,
0.9997, 0.9993, 0.9991,
        0.9995, 0.9995, 1.0000, 0.9999, 0.9995, 0.9995, 0.9994, 0.9998, 0.9996,
        1.0000, 0.9999, 1.0000, 0.9999, 0.9999, 0.9994, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([8.7056, 8.7897, 8.7636, 8.6463, 8.4558, 8.2076, 7.9147,
7.5873, 7.2337,
        6.8601, 6.4713, 6.0711, 5.6622, 5.2469, 4.8266, 4.4027, 3.9760, 3.5472,
        3.1169, 2.6854, 2.2531, 1.8200, 1.3865, 0.9526, 0.5185])
0,[1e-05,1][1, 10000, 1, 1],1682506150 saved
[1071125, 'tensor([0., 1., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.])',
225610.63312213228, 21262.174331621536, 0.07217718660831451, 1e-05, 1, 0,
'tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29,\n        29, 29, 29, 29, 29, 29,  0])', '[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.\n 1.]', '0,[1e-05,1][1, 10000, 1,
1],1682506150', 25, 50, 174060.25019204617, 225610.63312213225,
94497.47055325202, 131154.89666666667, 128142.59199999999, 65231.33608163523,
```

```
65242.24251310914, 80629.96159612676, 80629.96159612676, 109128.78816076377,
65242.24251310914, 80629.96159612676]
policy reset
----------------------------------------
iter  2  stage  24  ep  99999   adversary:  AdversaryModes.constant_132
  actions:  tensor([1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 1, 0, 0, 0, 1, 0, 0,
4, 0, 0, 0,
        0])
loss=  tensor(-0., grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.5700,
0.5700, 0.5700, 0.5700, 0.5700, 0.5700, 0.5700, 0.5700, 0.5700, 0.5700,
        0.5700, 0.5700, 0.5700, 0.5700, 0.5700, 0.5700, 0.5700, 0.5700, 0.5700,
        0.5700, 0.5700, 0.5700, 0.5700, 0.5700, 0.5700, 0.5700]) return=
139801.3687034911
probs of actions:  tensor([0.0586, 0.8760, 0.8988, 0.8761, 0.8584, 0.9093,
0.8702, 0.8950, 0.8991,
        0.8972, 0.9039, 0.0053, 0.8843, 0.0620, 0.8891, 0.8950, 0.8928, 0.0581,
        0.9003, 0.8979, 0.0050, 0.9160, 0.9029, 0.8696, 0.9885],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5111, 0.5274, 0.5361, 0.5426, 0.5476, 0.5513, 0.5541,
0.5562, 0.5578,
        0.5589, 0.5598, 0.5596, 0.5723, 0.5697, 0.5718, 0.5694, 0.5677, 0.5663,
        0.5692, 0.5675, 0.5647, 0.5805, 0.5759, 0.5726, 0.5700])
finalReturns:  tensor([0.])
----------------------------------------
iter  2  stage  23  ep  99999   adversary:  AdversaryModes.constant_132
  actions:  tensor([22,  0,  0, 15, 13,  0, 11,  0,  0,  0, 20,  5,  3,  0,  3,
0, 26, 25,
         0, 11,  0,  6,  2, 12,  0])
loss=  tensor(0.1463, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.2831,
1.2831, 1.2831, 1.2831, 1.2831, 1.2831, 1.2831, 1.2831, 1.2831, 1.2831,
        1.2831, 1.2831, 1.2831, 1.2831, 1.2831, 1.2831, 1.2831, 1.2831, 1.2831,
        1.2831, 1.2831, 1.2831, 1.2831, 1.2831, 1.2831, 0.6299]) return=
159276.00506941756
probs of actions:  tensor([0.0025, 0.2966, 0.3466, 0.0218, 0.0535, 0.3252,
0.0605, 0.3943, 0.3766,
        0.3660, 0.0257, 0.0161, 0.0400, 0.3518, 0.0500, 0.3720, 0.0043, 0.0210,
        0.4237, 0.0692, 0.2962, 0.0161, 0.0380, 0.0140, 0.9845],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.6065, 0.5953, 0.5645, 0.6225, 0.6719, 0.6316,
0.6670, 0.6401,
        0.6202, 0.5655, 0.6717, 0.6647, 0.6510, 0.6274, 0.6233, 0.5402, 0.6384,
        0.7707, 0.7035, 0.7216, 0.6763, 0.6735, 0.6388, 0.6784])
finalReturns:  tensor([0.0341, 0.0485])
----------------------------------------
iter  2  stage  22  ep  99999   adversary:  AdversaryModes.constant_132
  actions:  tensor([ 1, 23,  0, 24,  0, 23, 25,  0,  0,  1, 23, 23,  1, 23, 31,
20,  0, 25,
        23, 15, 20, 18, 23, 23,  0])
```

```
loss=  tensor(0.1068, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.3746,
2.3746, 2.3746, 2.3746, 2.3746, 2.3746, 2.3746, 2.3746, 2.3746,
        2.3746, 2.3746, 2.3746, 2.3746, 2.3746, 2.3746, 2.3746, 2.3746, 2.3746,
        2.3746, 2.3746, 2.3746, 2.3746, 2.3746, 1.5088, 0.7250]) return=
180250.60124621165
probs of actions:  tensor([0.0518, 0.3167, 0.1523, 0.0385, 0.0575, 0.4545,
0.1042, 0.2132, 0.0763,
        0.1119, 0.3796, 0.3270, 0.1043, 0.3741, 0.0090, 0.0816, 0.1816, 0.1185,
        0.2634, 0.0165, 0.0844, 0.0138, 0.7915, 0.7690, 0.9828],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5111, 0.4745, 0.6236, 0.5504, 0.6928, 0.6060, 0.6665,
0.7927, 0.7315,
        0.6871, 0.6060, 0.6760, 0.7837, 0.6764, 0.6880, 0.8233, 0.8730, 0.7265,
        0.7865, 0.8464, 0.8140, 0.8336, 0.8128, 0.8360, 0.9065])
finalReturns:  tensor([0.1808, 0.2337, 0.1815])
----------------------------------------
iter  2  stage  21  ep  99999   adversary:  AdversaryModes.constant_132
  actions:  tensor([31, 23, 25, 25, 25, 25, 23, 25, 23, 23, 23, 25, 25, 23, 25,
23, 25, 25,
        27, 25, 25, 25, 25, 25,  0])
loss=  tensor(0.7501, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.4017,
3.4017, 3.4017, 3.4017, 3.4017, 3.4017, 3.4017, 3.4017, 3.4017,
        3.4017, 3.4017, 3.4017, 3.4017, 3.4017, 3.4017, 3.4017, 3.4017, 3.4017,
        3.4017, 3.4017, 3.4017, 3.4017, 2.4026, 1.5243, 0.7315]) return=
213022.8652278334
probs of actions:  tensor([0.0111, 0.2492, 0.4587, 0.4976, 0.4984, 0.5060,
0.2525, 0.4668, 0.3060,
        0.2324, 0.2723, 0.5297, 0.4819, 0.2750, 0.5488, 0.2651, 0.5455, 0.4659,
        0.0139, 0.5334, 0.5165, 0.6597, 0.5791, 0.4736, 0.9465],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4151, 0.5891, 0.6531, 0.7198, 0.7717, 0.8117, 0.8520,
0.8561, 0.8857,
        0.8912, 0.8952, 0.8887, 0.9008, 0.9195, 0.9069, 0.9241, 0.9103, 0.9171,
        0.9118, 0.9360, 0.9363, 0.9366, 0.9369, 0.9370, 0.9996])
finalReturns:  tensor([0.4084, 0.4709, 0.4124, 0.2681])
----------------------------------------
iter  2  stage  20  ep  99999   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 25, 24, 25, 25, 31, 23, 25, 23, 24, 25, 29, 25, 25, 25,
25, 23, 25,
        25, 25, 25, 25, 25, 25,  0])
loss=  tensor(0.3533, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.0790,
4.0790, 4.0790, 4.0790, 4.0790, 4.0790, 4.0790, 4.0790, 4.0790,
        4.0790, 4.0790, 4.0790, 4.0790, 4.0790, 4.0790, 4.0790, 4.0790, 4.0790,
        4.0790, 4.0790, 4.0790, 3.0833, 2.2075, 1.4165, 0.6863]) return=
214816.26798917833
probs of actions:  tensor([0.0574, 0.6796, 0.0284, 0.7525, 0.7721, 0.0030,
0.1141, 0.6917, 0.1161,
        0.0350, 0.8065, 0.0533, 0.7493, 0.7699, 0.8023, 0.8069, 0.0802, 0.7284,
```

```
        0.7459, 0.8124, 0.8481, 0.9153, 0.8995, 0.8177, 0.9781],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5715, 0.6601, 0.7170, 0.7696, 0.7765, 0.8795,
0.8768, 0.9014,
        0.8982, 0.8994, 0.8872, 0.9358, 0.9363, 0.9366, 0.9368, 0.9466, 0.9271,
        0.9297, 0.9317, 0.9331, 0.9342, 0.9350, 0.9357, 0.9986])
finalReturns:  tensor([0.6577, 0.7202, 0.6618, 0.5178, 0.3124])
----------------------------------------
iter  2  stage  19  ep  99999   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 25, 25, 25, 25, 29, 29, 27, 29, 25, 29, 29, 25, 25,
29, 25, 25,
        25, 29, 25, 25, 25, 25,  0])
loss=  tensor(2.2123, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.7950,
4.7950, 4.7950, 4.7950, 4.7950, 4.7950, 4.7950, 4.7950, 4.7950,
        4.7950, 4.7950, 4.7950, 4.7950, 4.7950, 4.7950, 4.7950, 4.7950, 4.7950,
        4.7950, 4.7950, 3.7789, 2.8886, 2.0874, 1.3497, 0.6581]) return=
219146.74992654673
probs of actions:  tensor([0.3673, 0.4170, 0.5130, 0.5435, 0.5898, 0.5895,
0.4646, 0.4325, 0.0031,
        0.4868, 0.5454, 0.4070, 0.3855, 0.5566, 0.5777, 0.3509, 0.5618, 0.4698,
        0.4669, 0.3693, 0.5363, 0.7391, 0.8429, 0.8211, 0.9920],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6723, 0.7348, 0.7833, 0.8206, 0.8276,
0.8687, 0.9113,
        0.9140, 0.9562, 0.9299, 0.9466, 0.9808, 0.9699, 0.9401, 0.9759, 0.9663,
        0.9590, 0.9320, 0.9698, 0.9617, 0.9556, 0.9511, 1.0102])
finalReturns:  tensor([0.9853, 1.0694, 0.9899, 0.8295, 0.6115, 0.3521])
----------------------------------------
iter  2  stage  18  ep  99999   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 25,  0])
loss=  tensor(1.4374, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.6221,
5.6221, 5.6221, 5.6221, 5.6221, 5.6221, 5.6221, 5.6221, 5.6221,
        5.6221, 5.6221, 5.6221, 5.6221, 5.6221, 5.6221, 5.6221, 5.6221, 5.6221,
        5.6221, 4.5443, 3.6109, 2.7790, 2.0194, 1.3118, 0.6421]) return=
225619.69834256533
probs of actions:  tensor([0.9429, 0.9464, 0.9397, 0.9376, 0.9270, 0.9454,
0.9617, 0.9620, 0.9437,
        0.9725, 0.9606, 0.9671, 0.9557, 0.9405, 0.9506, 0.9565, 0.9554, 0.9611,
        0.9799, 0.9698, 0.9807, 0.9492, 0.8315, 0.2472, 0.9998],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 1.0182, 1.0602])
finalReturns:  tensor([1.4322, 1.5163, 1.4550, 1.2916, 1.0553, 0.7666, 0.4182])
----------------------------------------
```

```
iter  2  stage  17  ep  99999   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.1282, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.2397,
6.2397, 6.2397, 6.2397, 6.2397, 6.2397, 6.2397, 6.2397, 6.2397,
        6.2397, 6.2397, 6.2397, 6.2397, 6.2397, 6.2397, 6.2397, 6.2397, 6.2397,
        5.1632, 4.2307, 3.3994, 2.6402, 1.9330, 1.2634, 0.6215]) return=
225610.63312213228
probs of actions:  tensor([0.9887, 0.9903, 0.9890, 0.9885, 0.9857, 0.9906,
0.9939, 0.9941, 0.9898,
        0.9961, 0.9934, 0.9951, 0.9932, 0.9894, 0.9918, 0.9928, 0.9927, 0.9959,
        0.9982, 0.9969, 0.9978, 0.9934, 0.9720, 0.9220, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([1.8061, 1.8902, 1.8290, 1.6657, 1.4295, 1.1409, 0.8141,
0.4594])
----------------------------------------
iter  2  stage  16  ep  99999   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.2565, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.8407,
6.8407, 6.8407, 6.8407, 6.8407, 6.8407, 6.8407, 6.8407, 6.8407,
        6.8407, 6.8407, 6.8407, 6.8407, 6.8407, 6.8407, 6.8407, 5.7658,
        4.8345, 4.0040, 3.2454, 2.5386, 1.8694, 1.2278, 0.6064]) return=
225610.63312213228
probs of actions:  tensor([0.9848, 0.9873, 0.9857, 0.9847, 0.9815, 0.9870,
0.9919, 0.9923, 0.9860,
        0.9947, 0.9908, 0.9933, 0.9906, 0.9856, 0.9885, 0.9898, 0.9932, 0.9950,
        0.9983, 0.9976, 0.9975, 0.9915, 0.9593, 0.8347, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([2.1959, 2.2800, 2.2189, 2.0557, 1.8196, 1.5311, 1.2044,
0.8498, 0.4745])
----------------------------------------
iter  2  stage  15  ep  66414   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0475, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.4283,
7.4283, 7.4283, 7.4283, 7.4283, 7.4283, 7.4283, 7.4283, 7.4283,
```

```
        7.4283, 7.4283, 7.4283, 7.4283, 7.4283, 7.4283, 7.4283, 6.3558, 5.4260,
        4.5967, 3.8389, 3.1326, 2.4638, 1.8225, 1.2013, 0.5951]) return=
225610.63312213228
probs of actions:  tensor([0.9969, 0.9974, 0.9972, 0.9971, 0.9963, 0.9978,
0.9985, 0.9986, 0.9973,
        0.9991, 0.9984, 0.9989, 0.9985, 0.9973, 0.9980, 0.9990, 0.9994, 0.9998,
        1.0000, 0.9998, 1.0000, 0.9991, 0.9926, 0.9648, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([2.5967, 2.6808, 2.6198, 2.4567, 2.2208, 1.9324, 1.6059,
1.2513, 0.8762,
        0.4858])
----------------------------------------
iter  2  stage  14  ep  1469   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0459, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.0050,
8.0050, 8.0050, 8.0050, 8.0050, 8.0050, 8.0050, 8.0050, 8.0050,
        8.0050, 8.0050, 8.0050, 8.0050, 8.0050, 8.0050, 6.9354, 6.0078, 5.1799,
        4.4232, 3.7177, 3.0495, 2.4086, 1.7877, 1.1817, 0.5868]) return=
225610.63312213228
probs of actions:  tensor([0.9972, 0.9976, 0.9975, 0.9974, 0.9967, 0.9980,
0.9986, 0.9988, 0.9976,
        0.9992, 0.9986, 0.9990, 0.9986, 0.9976, 0.9990, 0.9991, 0.9994, 0.9998,
        1.0000, 0.9998, 1.0000, 0.9993, 0.9937, 0.9690, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([3.0055, 3.0896, 3.0288, 2.8659, 2.6302, 2.3420, 2.0156,
1.6611, 1.2861,
        0.8958, 0.4942])
----------------------------------------
iter  2  stage  13  ep  39701   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0311, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.5720,
8.5720, 8.5720, 8.5720, 8.5720, 8.5720, 8.5720, 8.5720, 8.5720,
        8.5720, 8.5720, 8.5720, 8.5720, 8.5720, 7.5064, 6.5815, 5.7556, 5.0003,
        4.2958, 3.6284, 2.9880, 2.3676, 1.7619, 1.1671, 0.5805]) return=
225610.63312213228
probs of actions:  tensor([0.9982, 0.9985, 0.9985, 0.9984, 0.9979, 0.9988,
```

```
0.9992, 0.9993, 0.9985,
        0.9995, 0.9992, 0.9994, 0.9992, 0.9990, 0.9994, 1.0000, 0.9999, 1.0000,
        1.0000, 1.0000, 1.0000, 0.9998, 0.9960, 0.9792, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([3.4201, 3.5042, 3.4436, 3.2810, 3.0455, 2.7576, 2.4313,
2.0771, 1.7022,
        1.3119, 0.9104, 0.5004])
---------------------------------------
iter  2  stage  12  ep  0   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0353, grad_fn=<NegBackward0>)   ,  base rewards= tensor([9.1296,
9.1296, 9.1296, 9.1296, 9.1296, 9.1296, 9.1296, 9.1296, 9.1296,
        9.1296, 9.1296, 9.1296, 9.1296, 8.0693, 7.1481, 6.3249, 5.5714, 4.8684,
        4.2019, 3.5622, 2.9423, 2.3370, 1.7426, 1.1562, 0.5759]) return=
225610.63312213228
probs of actions:  tensor([0.9982, 0.9985, 0.9985, 0.9984, 0.9979, 0.9988,
0.9992, 0.9993, 0.9985,
        0.9995, 0.9992, 0.9994, 0.9992, 0.9990, 0.9994, 1.0000, 0.9999, 1.0000,
        1.0000, 1.0000, 1.0000, 0.9998, 0.9960, 0.9792, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([3.8386, 3.9227, 3.8624, 3.7002, 3.4651, 3.1775, 2.8515,
2.4975, 2.1227,
        1.7327, 1.3312, 0.9213, 0.5051])
---------------------------------------
iter  2  stage  11  ep  0   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0389, grad_fn=<NegBackward0>)   ,  base rewards= tensor([9.6775,
9.6775, 9.6775, 9.6775, 9.6775, 9.6775, 9.6775, 9.6775, 9.6775,
        9.6775, 9.6775, 9.6775, 8.6243, 7.7080, 6.8883, 6.1374, 5.4361, 4.7710,
        4.1323, 3.5131, 2.9084, 2.3143, 1.7282, 1.1481, 0.5724]) return=
225610.63312213228
probs of actions:  tensor([0.9982, 0.9985, 0.9985, 0.9984, 0.9979, 0.9988,
0.9992, 0.9993, 0.9985,
        0.9995, 0.9992, 0.9994, 0.9992, 0.9990, 0.9994, 1.0000, 0.9999, 1.0000,
        1.0000, 1.0000, 1.0000, 0.9998, 0.9960, 0.9792, 1.0000],
       grad_fn=<ExpBackward0>)
```

```
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([4.2598, 4.3439, 4.2840, 4.1223, 3.8877, 3.6005, 3.2749,
2.9211, 2.5466,
        2.1567, 1.7554, 1.3456, 0.9294, 0.5085])
----------------------------------------
iter  2  stage  10  ep  0   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0440, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([10.2146, 10.2146, 10.2146, 10.2146, 10.2146, 10.2146, 10.2146, 10.2146,
        10.2146, 10.2146, 10.2146,  9.1707,  8.2611,  7.4460,  6.6984,  5.9995,
         5.3362,  4.6988,  4.0806,  3.4765,  2.8830,  2.2973,  1.7175,  1.1420,
         0.5698]) return=  225610.63312213228
probs of actions:  tensor([0.9982, 0.9985, 0.9985, 0.9984, 0.9980, 0.9988,
0.9992, 0.9993, 0.9985,
        0.9995, 0.9992, 0.9994, 0.9992, 0.9990, 0.9994, 1.0000, 0.9999, 1.0000,
        1.0000, 1.0000, 1.0000, 0.9998, 0.9961, 0.9792, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([4.6825, 4.7666, 4.7072, 4.5461, 4.3122, 4.0255, 3.7004,
3.3470, 2.9728,
        2.5832, 2.1820, 1.7724, 1.3563, 0.9355, 0.5111])
----------------------------------------
iter  2  stage  9  ep  0   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0476, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([10.7388, 10.7388, 10.7388, 10.7388, 10.7388, 10.7388, 10.7388, 10.7388,
        10.7388, 10.7388,  9.7074,  8.8064,  7.9975,  7.2544,  6.5588,  5.8977,
         5.2621,  4.6452,  4.0421,  3.4493,  2.8641,  2.2847,  1.7095,  1.1375,
         0.5678]) return=  225610.63312213228
probs of actions:  tensor([0.9982, 0.9985, 0.9985, 0.9984, 0.9980, 0.9988,
0.9992, 0.9993, 0.9985,
        0.9995, 0.9992, 0.9994, 0.9992, 0.9990, 0.9994, 1.0000, 0.9999, 1.0000,
        1.0000, 1.0000, 1.0000, 0.9998, 0.9961, 0.9792, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
```

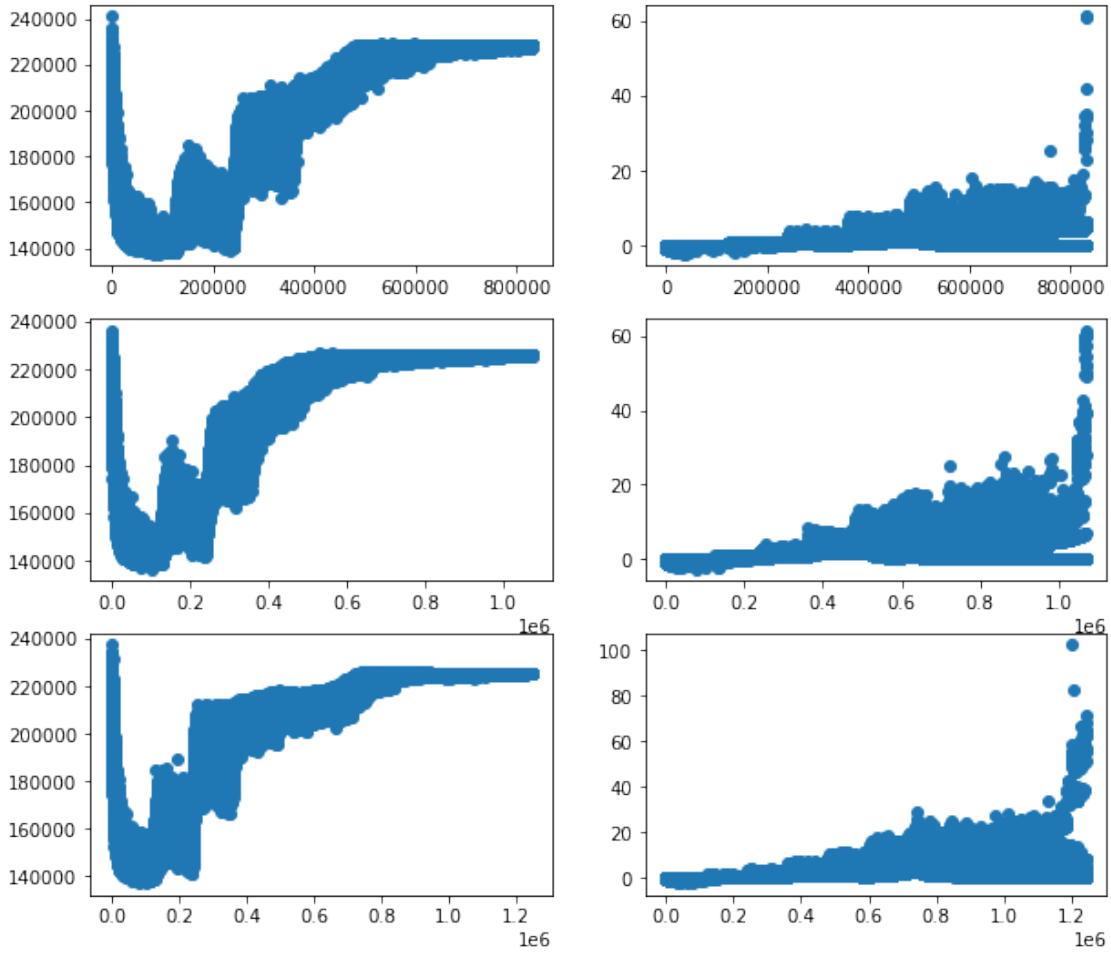
```
finalReturns:  tensor([5.1056, 5.1897, 5.1309, 4.9707, 4.7377, 4.4518, 4.1273,
3.7745, 3.4007,
        3.0113, 2.6104, 2.2010, 1.7851, 1.3643, 0.9401, 0.5131])
----------------------------------------
iter  2  stage  8  ep  12279   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0442, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([11.2472, 11.2472, 11.2472, 11.2472, 11.2472, 11.2472, 11.2472, 11.2472,
        11.2472, 10.2322,  9.3428,  8.5420,  7.8048,  7.1135,  6.4556,  5.8223,
         5.2071,  4.6052,  4.0134,  3.4289,  2.8500,  2.2752,  1.7035,  1.1341,
         0.5664]) return=  225610.63312213228
probs of actions:  tensor([0.9982, 0.9985, 0.9984, 0.9984, 0.9979, 0.9988,
0.9991, 0.9993, 0.9990,
        1.0000, 1.0000, 0.9997, 0.9994, 0.9990, 0.9994, 1.0000, 0.9999, 1.0000,
        1.0000, 1.0000, 1.0000, 0.9998, 0.9958, 0.9800, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([5.5282, 5.6123, 5.5544, 5.3953, 5.1634, 4.8785, 4.5549,
4.2028, 3.8295,
        3.4406, 3.0400, 2.6308, 2.2151, 1.7945, 1.3704, 0.9435, 0.5145])
----------------------------------------
iter  2  stage  7  ep  0   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0494, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([11.7354, 11.7354, 11.7354, 11.7354, 11.7354, 11.7354, 11.7354, 11.7354,
        10.7421,  9.8680,  9.0781,  8.3488,  7.6631,  7.0094,  6.3791,  5.7662,
         5.1660,  4.5754,  3.9919,  3.4137,  2.8394,  2.2681,  1.6990,  1.1315,
         0.5653]) return=  225610.63312213228
probs of actions:  tensor([0.9982, 0.9985, 0.9984, 0.9984, 0.9979, 0.9988,
0.9991, 0.9993, 0.9990,
        1.0000, 1.0000, 0.9997, 0.9994, 0.9990, 0.9994, 1.0000, 0.9999, 1.0000,
        1.0000, 1.0000, 1.0000, 0.9998, 0.9958, 0.9800, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([5.9491, 6.0332, 5.9765, 5.8190, 5.5886, 5.3051, 4.9826,
4.6314, 4.2589,
        3.8705, 3.4704, 3.0615, 2.6460, 2.2257, 1.8016, 1.3749, 0.9460, 0.5156])
----------------------------------------
```

```
iter  2  stage  6  ep  0   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0560, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([12.1977, 12.1977, 12.1977, 12.1977, 12.1977, 12.1977, 12.1977, 11.2331,
        10.3791,  9.6037,  8.8847,  8.2066,  7.5584,  6.9322,  6.3223,  5.7244,
         5.1354,  4.5532,  3.9759,  3.4023,  2.8315,  2.2627,  1.6956,  1.1296,
         0.5645]) return=  225610.63312213228
probs of actions:  tensor([0.9982, 0.9985, 0.9984, 0.9984, 0.9979, 0.9988,
0.9991, 0.9993, 0.9990,
        1.0000, 1.0000, 0.9997, 0.9994, 0.9990, 0.9994, 1.0000, 0.9999, 1.0000,
        1.0000, 1.0000, 1.0000, 0.9998, 0.9958, 0.9800, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([6.3673, 6.4514, 6.3962, 6.2408, 6.0125, 5.7308, 5.4098,
5.0598, 4.6882,
        4.3007, 3.9011, 3.4927, 3.0775, 2.6574, 2.2336, 1.8070, 1.3782, 0.9479,
        0.5165])
----------------------------------------
iter  2  stage  5  ep  90   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0596, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([12.6268, 12.6268, 12.6268, 12.6268, 12.6268, 12.6268, 11.6997, 10.8722,
        10.1157,  9.4104,  8.7424,  8.1016,  7.4808,  6.8749,  6.2799,  5.6932,
         5.1126,  4.5365,  3.9639,  3.3937,  2.8255,  2.2588,  1.6931,  1.1282,
         0.5639]) return=  225610.63312213228
probs of actions:  tensor([0.9983, 0.9985, 0.9985, 0.9985, 0.9980, 0.9990,
0.9993, 0.9994, 0.9991,
        1.0000, 1.0000, 0.9997, 0.9994, 0.9991, 0.9994, 1.0000, 0.9999, 1.0000,
        1.0000, 1.0000, 1.0000, 0.9998, 0.9960, 0.9811, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([6.7813, 6.8654, 6.8123, 6.6596, 6.4340, 6.1547, 5.8358,
5.4874, 5.1172,
        4.7306, 4.3318, 3.9240, 3.5093, 3.0895, 2.6660, 2.2395, 1.8109, 1.3808,
        0.9494, 0.5171])
----------------------------------------
iter  2  stage  4  ep  882   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
```

```
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0599, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([13.0129, 13.0129, 13.0129, 13.0129, 13.0129, 12.1346, 11.3419, 10.6103,
         9.9232,  9.2683,  8.6372,  8.0237,  7.4231,  6.8321,  6.2483,  5.6699,
         5.0955,  4.5241,  3.9549,  3.3874,  2.8211,  2.2558,  1.6912,  1.1271,
         0.5634]) return=  225610.63312213228
probs of actions:  tensor([0.9984, 0.9987, 0.9986, 0.9986, 0.9990, 0.9991,
0.9995, 0.9995, 0.9992,
        1.0000, 1.0000, 0.9997, 0.9995, 0.9992, 0.9995, 1.0000, 0.9999, 1.0000,
        1.0000, 1.0000, 1.0000, 0.9998, 0.9963, 0.9830, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([7.1893, 7.2734, 7.2232, 7.0741, 6.8522, 6.5761, 6.2599,
5.9137, 5.5452,
        5.1599, 4.7622, 4.3551, 3.9411, 3.5218, 3.0985, 2.6724, 2.2440, 1.8139,
        1.3826, 0.9504, 0.5175])
----------------------------------------
iter  2  stage  3  ep  26351   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0595, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([13.3441, 13.3441, 13.3441, 13.3441, 12.5289, 11.7813, 11.0824, 10.4190,
         9.7816,  9.1633,  8.5593,  7.9657,  7.3800,  6.8002,  6.2247,  5.6525,
         5.0827,  4.5147,  3.9481,  3.3826,  2.8178,  2.2535,  1.6897,  1.1263,
         0.5630]) return=  225610.63312213228
probs of actions:  tensor([0.9984, 0.9986, 0.9986, 0.9990, 0.9992, 0.9989,
0.9997, 0.9997, 0.9993,
        1.0000, 1.0000, 1.0000, 0.9998, 0.9993, 0.9996, 1.0000, 0.9999, 1.0000,
        1.0000, 1.0000, 1.0000, 1.0000, 0.9961, 0.9837, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([7.5892, 7.6733, 7.6268, 7.4827, 7.2655, 6.9938, 6.6811,
6.3379, 5.9716,
        5.5882, 5.1918, 4.7858, 4.3726, 3.9539, 3.5311, 3.1053, 2.6771, 2.2473,
        1.8162, 1.3841, 0.9512, 0.5179])
----------------------------------------
iter  2  stage  2  ep  289   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 25,  0])
```

```
loss=  tensor(3.9982, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([13.6060, 13.6060, 13.6060, 12.8713, 12.1817, 11.5251, 10.8928, 10.2782,
         9.6769,  9.0855,  8.5013,  7.9226,  7.3479,  6.7763,  6.2070,  5.6394,
         5.0731,  4.5077,  3.9431,  3.3790,  2.8153,  2.2519,  1.6887,  1.1257,
         0.5628]) return=  225619.69834256533
probs of actions:  tensor([0.9984, 0.9987, 0.9990, 0.9990, 0.9992, 0.9989,
0.9997, 0.9997, 0.9993,
        1.0000, 1.0000, 1.0000, 0.9998, 0.9994, 0.9996, 1.0000, 0.9999, 1.0000,
        1.0000, 1.0000, 1.0000, 1.0000, 0.9961, 0.0159, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 1.0182, 1.0602])
finalReturns:  tensor([7.9789, 8.0630, 8.0214, 7.8838, 7.6732, 7.4072, 7.0993,
6.7599, 6.3967,
        6.0157, 5.6212, 5.2166, 4.8044, 4.3865, 3.9644, 3.5390, 3.1112, 2.6816,
        2.2507, 1.8188, 1.3860, 0.9528, 0.4974])
----------------------------------------
iter  2  stage  1  ep  7933   adversary:  AdversaryModes.constant_132
  actions:  tensor([25, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0748, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([13.7203, 13.7203, 13.1021, 12.4981, 11.9046, 11.3189, 10.7391, 10.1636,
         9.5914,  9.0216,  8.4536,  7.8870,  7.3215,  6.7567,  6.1925,  5.6287,
         5.0652,  4.5020,  3.9389,  3.3760,  2.8132,  2.2505,  1.6878,  1.1252,
         0.5626]) return=  225109.1821149025
probs of actions:  tensor([0.0017, 0.9990, 0.9990, 0.9991, 0.9993, 0.9989,
0.9997, 0.9999, 0.9993,
        1.0000, 1.0000, 1.0000, 0.9999, 0.9995, 0.9996, 1.0000, 1.0000, 1.0000,
        1.0000, 1.0000, 1.0000, 1.0000, 0.9959, 0.9832, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5341, 0.6379, 0.7210, 0.7863, 0.8369, 0.8759,
0.9056, 0.9282,
        0.9453, 0.9582, 0.9680, 0.9753, 0.9808, 0.9850, 0.9881, 0.9905, 0.9922,
        0.9935, 0.9945, 0.9953, 0.9958, 0.9962, 0.9966, 1.0809])
finalReturns:  tensor([8.3419, 8.4260, 8.3922, 8.2647, 8.0641, 7.8069, 7.5065,
7.1732, 6.8147,
        6.4374, 6.0458, 5.6433, 5.2328, 4.8162, 4.3950, 3.9703, 3.5431, 3.1139,
        2.6833, 2.2516, 1.8191, 1.3859, 0.9523, 0.5183])
----------------------------------------
iter  2  stage  0  ep  8878   adversary:  AdversaryModes.constant_132
  actions:  tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29, 29, 29,
        29, 29, 29, 29, 29, 29,  0])
loss=  tensor(0.0617, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([13.8555, 13.3442, 12.8204, 12.2871, 11.7465, 11.2005, 10.6504, 10.0972,
```

```
         9.5417,  8.9844,  8.4258,  7.8663,  7.3060,  6.7452,  6.1839,  5.6224,
         5.0606,  4.4986,  3.9365,  3.3743,  2.8120,  2.2497,  1.6873,  1.1249,
         0.5624]) return=  225610.63312213228
probs of actions:  tensor([0.9990, 0.9992, 0.9993, 0.9995, 0.9995, 0.9992,
0.9998, 0.9999, 0.9996,
        1.0000, 1.0000, 1.0000, 0.9999, 0.9999, 0.9997, 1.0000, 1.0000, 1.0000,
        1.0000, 1.0000, 1.0000, 1.0000, 0.9972, 0.9874, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4271, 0.5499, 0.6507, 0.7311, 0.7942, 0.8430, 0.8805,
0.9091, 0.9309,
        0.9473, 0.9598, 0.9691, 0.9762, 0.9815, 0.9855, 0.9885, 0.9907, 0.9924,
        0.9937, 0.9946, 0.9954, 0.9959, 0.9963, 0.9966, 1.0809])
finalReturns:  tensor([8.7056, 8.7897, 8.7636, 8.6463, 8.4558, 8.2076, 7.9147,
7.5873, 7.2337,
        6.8601, 6.4713, 6.0711, 5.6622, 5.2469, 4.8266, 4.4027, 3.9760, 3.5472,
        3.1169, 2.6854, 2.2531, 1.8200, 1.3865, 0.9526, 0.5185])
0,[1e-05,1][1, 10000, 1, 1],1682631035 saved
[1244302, 'tensor([0., 1., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.])',
225610.63312213228, 21262.174331621536, 0.061683181673288345, 1e-05, 1, 0,
'tensor([29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29,
29,\n        29, 29, 29, 29, 29, 29,  0])', '[1.   1.   1.   1.   1.   1.   1.
1.   1.   1.   1.   1.   1.   1.   1.\n 1.   1.   1.   1.   1.   1.   1.   1.
0.99 1.  ]', '0,[1e-05,1][1, 10000, 1, 1],1682631035', 25, 50,
174060.25019204617, 225611.5396441756, 94497.47055325202, 131149.86666666667,
128142.59199999999, 65242.24251310914, 65242.24251310914, 80629.96159612676,
80629.96159612676, 109136.08826110291, 65218.96207421804, 80629.96159612676]
```




```
policy reset
----------------------------------------
iter  0  stage  24  ep  99999   adversary:  AdversaryModes.constant_95
  actions:  tensor([ 0,  0, 38,  0,  1,  0,  2,  0,  0,  1,  0,  0,  1,  0,  0,
0,  0,  0,
         1,  0,  1,  0,  0,  1,  0])
loss=  tensor(-0., grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.1482,
0.1482, 0.1482, 0.1482, 0.1482, 0.1482, 0.1482, 0.1482, 0.1482,
        0.1482, 0.1482, 0.1482, 0.1482, 0.1482, 0.1482, 0.1482, 0.1482, 0.1482,
        0.1482, 0.1482, 0.1482, 0.1482, 0.1482, 0.1482, 0.1482]) return=
51639.70510999265
probs of actions:  tensor([8.0034e-01, 8.2695e-01, 3.4396e-04, 8.0664e-01,
1.2781e-01, 7.9923e-01,
        1.9496e-02, 7.3233e-01, 7.9346e-01, 1.3482e-01, 7.8419e-01, 7.5348e-01,
        1.3699e-01, 7.8668e-01, 7.3267e-01, 7.2963e-01, 7.8819e-01, 7.6218e-01,
        1.3528e-01, 7.8927e-01, 1.3168e-01, 7.7874e-01, 7.6106e-01, 1.3310e-01,
        9.4036e-01], grad_fn=<ExpBackward0>)
```

```
rewards:  tensor([0.5112, 0.3985, 0.1787, 0.3799, 0.3104, 0.2657, 0.2315,
0.2127, 0.1944,
        0.1811, 0.1737, 0.1661, 0.1604, 0.1584, 0.1548, 0.1522, 0.1502, 0.1488,
        0.1476, 0.1488, 0.1476, 0.1488, 0.1477, 0.1468, 0.1482])
finalReturns:  tensor([0.])
----------------------------------------
iter  0  stage  23  ep  99999   adversary:  AdversaryModes.constant_95
  actions:  tensor([ 6,  4,  8,  0,  6,  6,  7,  0, 13,  3,  7,  0,  5,  6,  7,
7,  4,  4,
         0,  7,  6,  0,  1,  4,  0])
loss=  tensor(0.0107, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.3310,
0.3310, 0.3310, 0.3310, 0.3310, 0.3310, 0.3310, 0.3310, 0.3310,
        0.3310, 0.3310, 0.3310, 0.3310, 0.3310, 0.3310, 0.3310, 0.3310, 0.3310,
        0.3310, 0.3310, 0.3310, 0.3310, 0.3310, 0.3310, 0.1624]) return=
56826.748788899706
probs of actions:  tensor([0.2833, 0.2137, 0.0058, 0.1687, 0.2598, 0.2771,
0.2007, 0.1315, 0.0016,
        0.0535, 0.2105, 0.1792, 0.0410, 0.2486, 0.2121, 0.2264, 0.1918, 0.1917,
        0.1469, 0.1903, 0.2626, 0.1697, 0.0305, 0.1974, 0.9888],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5076, 0.4160, 0.3413, 0.3105, 0.2595, 0.2411, 0.2265,
0.2240, 0.1856,
        0.2153, 0.1987, 0.2033, 0.1852, 0.1834, 0.1837, 0.1871, 0.1930, 0.1884,
        0.1865, 0.1706, 0.1785, 0.1849, 0.1742, 0.1670, 0.1706])
finalReturns:  tensor([0.0066, 0.0082])
----------------------------------------
iter  0  stage  22  ep  99999   adversary:  AdversaryModes.constant_95
  actions:  tensor([ 7,  7,  7,  7,  7, 14,  7,  7,  7,  7, 12,  7,  7,  6,  7,
7,  7,  7,
         7,  7,  7,  7,  7,  7,  0])
loss=  tensor(0.0059, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.5673,
0.5673, 0.5673, 0.5673, 0.5673, 0.5673, 0.5673, 0.5673, 0.5673,
        0.5673, 0.5673, 0.5673, 0.5673, 0.5673, 0.5673, 0.5673, 0.5673, 0.5673,
        0.5673, 0.5673, 0.5673, 0.5673, 0.5673, 0.3640, 0.1763]) return=
61396.95122420932
probs of actions:  tensor([0.9033, 0.9047, 0.9016, 0.8795, 0.8839, 0.0039,
0.8793, 0.8599, 0.8975,
        0.8653, 0.0223, 0.8265, 0.8999, 0.0598, 0.8802, 0.8758, 0.8795, 0.8736,
        0.8427, 0.8736, 0.8772, 0.8745, 0.9073, 0.9292, 0.9986],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5063, 0.4160, 0.3540, 0.3107, 0.2801, 0.2435, 0.2600,
0.2436, 0.2317,
        0.2229, 0.2069, 0.2234, 0.2168, 0.2132, 0.2060, 0.2039, 0.2023, 0.2011,
        0.2002, 0.1996, 0.1991, 0.1987, 0.1984, 0.1982, 0.2030])
finalReturns:  tensor([0.0323, 0.0372, 0.0267])
----------------------------------------
iter  0  stage  21  ep  99999   adversary:  AdversaryModes.constant_95
  actions:  tensor([12, 12, 12, 12, 12,  9, 12, 12, 12, 12, 12,  7, 12, 12,  7,
```

```
7, 12, 12,
        10, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0066, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.8348,
0.8348, 0.8348, 0.8348, 0.8348, 0.8348, 0.8348, 0.8348, 0.8348,
        0.8348, 0.8348, 0.8348, 0.8348, 0.8348, 0.8348, 0.8348, 0.8348, 0.8348,
        0.8348, 0.8348, 0.8348, 0.8348, 0.5933, 0.3783, 0.1824]) return=
67207.34217323098
probs of actions:  tensor([0.9530, 0.9572, 0.9566, 0.9385, 0.9367, 0.0018,
0.9516, 0.9293, 0.9447,
        0.9415, 0.9464, 0.0281, 0.9350, 0.9386, 0.0333, 0.0270, 0.9371, 0.9286,
        0.0046, 0.9552, 0.9292, 0.9849, 0.9793, 0.9724, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3345, 0.3083, 0.2955, 0.2673,
0.2592, 0.2532,
        0.2487, 0.2454, 0.2525, 0.2286, 0.2304, 0.2412, 0.2299, 0.2121, 0.2179,
        0.2266, 0.2207, 0.2244, 0.2271, 0.2292, 0.2308, 0.2464])
finalReturns:  tensor([0.0988, 0.1132, 0.0989, 0.0641])
----------------------------------------
iter  0  stage  20  ep  99999   adversary:  AdversaryModes.constant_95
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0034, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.0332,
1.0332, 1.0332, 1.0332, 1.0332, 1.0332, 1.0332, 1.0332, 1.0332,
        1.0332, 1.0332, 1.0332, 1.0332, 1.0332, 1.0332, 1.0332, 1.0332, 1.0332,
        1.0332, 1.0332, 1.0332, 0.7825, 0.5612, 0.3606, 0.1749]) return=
68694.09895647192
probs of actions:  tensor([0.9908, 0.9920, 0.9920, 0.9888, 0.9875, 0.9919,
0.9900, 0.9860, 0.9891,
        0.9871, 0.9900, 0.9854, 0.9871, 0.9887, 0.9851, 0.9900, 0.9875, 0.9859,
        0.9865, 0.9914, 0.9903, 0.9968, 0.9954, 0.9957, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3345, 0.3083, 0.2892, 0.2753,
0.2651, 0.2576,
        0.2520, 0.2479, 0.2448, 0.2425, 0.2407, 0.2394, 0.2385, 0.2378, 0.2372,
        0.2368, 0.2365, 0.2363, 0.2361, 0.2360, 0.2359, 0.2502])
finalReturns:  tensor([0.1613, 0.1757, 0.1609, 0.1255, 0.0753])
----------------------------------------
iter  0  stage  19  ep  99999   adversary:  AdversaryModes.constant_95
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0025, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.2009,
1.2009, 1.2009, 1.2009, 1.2009, 1.2009, 1.2009, 1.2009, 1.2009,
        1.2009, 1.2009, 1.2009, 1.2009, 1.2009, 1.2009, 1.2009, 1.2009, 1.2009,
        1.2009, 1.2009, 1.2009, 0.9500, 0.7285, 0.5277, 0.3420, 0.1670]) return=
68694.09895647192
probs of actions:  tensor([0.9955, 0.9963, 0.9964, 0.9952, 0.9943, 0.9963,
```

```
0.9954, 0.9935, 0.9951,
        0.9938, 0.9956, 0.9930, 0.9943, 0.9949, 0.9932, 0.9956, 0.9942, 0.9936,
        0.9938, 0.9969, 0.9960, 0.9991, 0.9979, 0.9984, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3345, 0.3083, 0.2892, 0.2753,
0.2651, 0.2576,
        0.2520, 0.2479, 0.2448, 0.2425, 0.2407, 0.2394, 0.2385, 0.2378, 0.2372,
        0.2368, 0.2365, 0.2363, 0.2361, 0.2360, 0.2359, 0.2502])
finalReturns:  tensor([0.2301, 0.2445, 0.2297, 0.1944, 0.1441, 0.0832])
----------------------------------------
iter  0  stage  18  ep  99999   adversary:  AdversaryModes.constant_95
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0020, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.3631,
1.3631, 1.3631, 1.3631, 1.3631, 1.3631, 1.3631, 1.3631, 1.3631,
        1.3631, 1.3631, 1.3631, 1.3631, 1.3631, 1.3631, 1.3631, 1.3631, 1.3631,
        1.3631, 1.1119, 0.8901, 0.6892, 0.5034, 0.3283, 0.1613]) return=
68694.09895647192
probs of actions:  tensor([0.9976, 0.9981, 0.9982, 0.9974, 0.9970, 0.9981,
0.9975, 0.9964, 0.9973,
        0.9966, 0.9977, 0.9961, 0.9969, 0.9973, 0.9962, 0.9976, 0.9969, 0.9965,
        0.9972, 0.9989, 0.9987, 0.9997, 0.9990, 0.9994, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3345, 0.3083, 0.2892, 0.2753,
0.2651, 0.2576,
        0.2520, 0.2479, 0.2448, 0.2425, 0.2407, 0.2394, 0.2385, 0.2378, 0.2372,
        0.2368, 0.2365, 0.2363, 0.2361, 0.2360, 0.2359, 0.2502])
finalReturns:  tensor([0.3047, 0.3191, 0.3044, 0.2690, 0.2187, 0.1578, 0.0890])
----------------------------------------
iter  0  stage  17  ep  45969   adversary:  AdversaryModes.constant_95
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0007, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.5214,
1.5214, 1.5214, 1.5214, 1.5214, 1.5214, 1.5214, 1.5214, 1.5214,
        1.5214, 1.5214, 1.5214, 1.5214, 1.5214, 1.5214, 1.5214, 1.5214, 1.5214,
        1.2697, 1.0477, 0.8466, 0.6606, 0.4855, 0.3183, 0.1570]) return=
68694.09895647192
probs of actions:  tensor([0.9993, 0.9994, 0.9995, 0.9992, 0.9991, 0.9994,
0.9992, 0.9988, 0.9991,
        0.9989, 0.9993, 0.9987, 0.9990, 0.9992, 0.9989, 0.9992, 0.9991, 0.9990,
        0.9998, 0.9998, 0.9998, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3345, 0.3083, 0.2892, 0.2753,
0.2651, 0.2576,
        0.2520, 0.2479, 0.2448, 0.2425, 0.2407, 0.2394, 0.2385, 0.2378, 0.2372,
        0.2368, 0.2365, 0.2363, 0.2361, 0.2360, 0.2359, 0.2502])
```

```
finalReturns:  tensor([0.3837, 0.3981, 0.3833, 0.3478, 0.2976, 0.2366, 0.1678,
0.0932])
----------------------------------------
iter  0  stage  16  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0012, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.6770,
1.6770, 1.6770, 1.6770, 1.6770, 1.6770, 1.6770, 1.6770, 1.6770, 1.6770,
        1.6770, 1.6770, 1.6770, 1.6770, 1.6770, 1.6770, 1.6770, 1.6770, 1.4248,
        1.2024, 1.0011, 0.8148, 0.6396, 0.4723, 0.3109, 0.1538]) return=
68694.09895647192
probs of actions:  tensor([0.9993, 0.9994, 0.9995, 0.9992, 0.9991, 0.9994,
0.9992, 0.9988, 0.9991,
        0.9989, 0.9993, 0.9987, 0.9990, 0.9992, 0.9989, 0.9992, 0.9991, 0.9990,
        0.9998, 0.9998, 0.9998, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3345, 0.3083, 0.2892, 0.2753,
0.2651, 0.2576,
        0.2520, 0.2479, 0.2448, 0.2425, 0.2407, 0.2394, 0.2385, 0.2378, 0.2372,
        0.2368, 0.2365, 0.2363, 0.2361, 0.2360, 0.2359, 0.2502])
finalReturns:  tensor([0.4658, 0.4802, 0.4654, 0.4299, 0.3796, 0.3186, 0.2498,
0.1752, 0.0964])
----------------------------------------
iter  0  stage  15  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0019, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.8309,
1.8309, 1.8309, 1.8309, 1.8309, 1.8309, 1.8309, 1.8309, 1.8309, 1.8309,
        1.8309, 1.8309, 1.8309, 1.8309, 1.8309, 1.8309, 1.8309, 1.8309, 1.5780,
        1.3551, 1.1534, 0.9669, 0.7914, 0.6240, 0.4625, 0.3054, 0.1515]) return=
68694.09895647192
probs of actions:  tensor([0.9993, 0.9994, 0.9995, 0.9992, 0.9991, 0.9994,
0.9992, 0.9988, 0.9991,
        0.9989, 0.9993, 0.9987, 0.9990, 0.9992, 0.9989, 0.9992, 0.9991, 0.9990,
        0.9998, 0.9998, 0.9998, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3345, 0.3083, 0.2892, 0.2753,
0.2651, 0.2576,
        0.2520, 0.2479, 0.2448, 0.2425, 0.2407, 0.2394, 0.2385, 0.2378, 0.2372,
        0.2368, 0.2365, 0.2363, 0.2361, 0.2360, 0.2359, 0.2502])
finalReturns:  tensor([0.5504, 0.5648, 0.5499, 0.5144, 0.4641, 0.4030, 0.3342,
0.2596, 0.1807,
        0.0987])
----------------------------------------
iter  0  stage  14  ep  66   adversary:  AdversaryModes.constant_95
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
```

```
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0025, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.9839,
1.9839, 1.9839, 1.9839, 1.9839, 1.9839, 1.9839, 1.9839, 1.9839,
        1.9839, 1.9839, 1.9839, 1.9839, 1.9839, 1.9839, 1.7301, 1.5065, 1.3043,
        1.1175, 0.9417, 0.7741, 0.6125, 0.4553, 0.3013, 0.1497]) return=
68694.09895647192
probs of actions:  tensor([0.9993, 0.9995, 0.9995, 0.9992, 0.9991, 0.9994,
0.9992, 0.9989, 0.9992,
        0.9989, 0.9993, 0.9987, 0.9991, 0.9992, 0.9990, 0.9993, 0.9992, 0.9990,
        0.9998, 0.9998, 0.9998, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3345, 0.3083, 0.2892, 0.2753,
0.2651, 0.2576,
        0.2520, 0.2479, 0.2448, 0.2425, 0.2407, 0.2394, 0.2385, 0.2378, 0.2372,
        0.2368, 0.2365, 0.2363, 0.2361, 0.2360, 0.2359, 0.2502])
finalReturns:  tensor([0.6368, 0.6512, 0.6363, 0.6007, 0.5503, 0.4892, 0.4203,
0.3457, 0.2668,
        0.1848, 0.1005])
----------------------------------------
iter  0  stage  13  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0034, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.1368,
2.1368, 2.1368, 2.1368, 2.1368, 2.1368, 2.1368, 2.1368, 2.1368,
        2.1368, 2.1368, 2.1368, 2.1368, 2.1368, 1.8817, 1.6572, 1.4543, 1.2670,
        1.0910, 0.9231, 0.7613, 0.6040, 0.4499, 0.2982, 0.1484]) return=
68694.09895647192
probs of actions:  tensor([0.9993, 0.9995, 0.9995, 0.9992, 0.9991, 0.9994,
0.9992, 0.9989, 0.9992,
        0.9989, 0.9993, 0.9987, 0.9991, 0.9992, 0.9990, 0.9993, 0.9992, 0.9990,
        0.9998, 0.9998, 0.9998, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3345, 0.3083, 0.2892, 0.2753,
0.2651, 0.2576,
        0.2520, 0.2479, 0.2448, 0.2425, 0.2407, 0.2394, 0.2385, 0.2378, 0.2372,
        0.2368, 0.2365, 0.2363, 0.2361, 0.2360, 0.2359, 0.2502])
finalReturns:  tensor([0.7246, 0.7390, 0.7241, 0.6884, 0.6380, 0.5768, 0.5079,
0.4332, 0.3542,
        0.2722, 0.1879, 0.1018])
----------------------------------------
iter  0  stage  12  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0044, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.2903,
2.2903, 2.2903, 2.2903, 2.2903, 2.2903, 2.2903, 2.2903, 2.2903,
```

```
        2.2903, 2.2903, 2.2903, 2.2903, 2.0334, 1.8077, 1.6040, 1.4161, 1.2396,
        1.0714, 0.9093, 0.7518, 0.5976, 0.4458, 0.2960, 0.1475]) return=
68694.09895647192
probs of actions:  tensor([0.9993, 0.9995, 0.9995, 0.9992, 0.9991, 0.9994,
0.9992, 0.9989, 0.9992,
        0.9989, 0.9993, 0.9987, 0.9991, 0.9992, 0.9990, 0.9993, 0.9992, 0.9990,
        0.9998, 0.9998, 0.9998, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3345, 0.3083, 0.2892, 0.2753,
0.2651, 0.2576,
        0.2520, 0.2479, 0.2448, 0.2425, 0.2407, 0.2394, 0.2385, 0.2378, 0.2372,
        0.2368, 0.2365, 0.2363, 0.2361, 0.2360, 0.2359, 0.2502])
finalReturns:  tensor([0.8136, 0.8280, 0.8130, 0.7773, 0.7267, 0.6655, 0.5964,
0.5217, 0.4427,
        0.3606, 0.2762, 0.1901, 0.1027])
----------------------------------------
iter  0  stage  11  ep  5857   adversary:  AdversaryModes.constant_95
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0029, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.4451,
2.4451, 2.4451, 2.4451, 2.4451, 2.4451, 2.4451, 2.4451, 2.4451,
        2.4451, 2.4451, 2.4451, 2.1859, 1.9586, 1.7537, 1.5650, 1.3879, 1.2192,
        1.0569, 0.8991, 0.7447, 0.5928, 0.4428, 0.2943, 0.1467]) return=
68694.09895647192
probs of actions:  tensor([0.9995, 0.9996, 0.9996, 0.9994, 0.9993, 0.9996,
0.9994, 0.9992, 0.9994,
        0.9991, 0.9995, 0.9990, 0.9998, 0.9997, 0.9995, 0.9998, 0.9997, 0.9993,
        0.9998, 0.9999, 0.9999, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3345, 0.3083, 0.2892, 0.2753,
0.2651, 0.2576,
        0.2520, 0.2479, 0.2448, 0.2425, 0.2407, 0.2394, 0.2385, 0.2378, 0.2372,
        0.2368, 0.2365, 0.2363, 0.2361, 0.2360, 0.2359, 0.2502])
finalReturns:  tensor([0.9036, 0.9180, 0.9029, 0.8670, 0.8163, 0.7549, 0.6858,
0.6110, 0.5319,
        0.4498, 0.3654, 0.2792, 0.1918, 0.1035])
----------------------------------------
iter  0  stage  10  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0036, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.6021,
2.6021, 2.6021, 2.6021, 2.6021, 2.6021, 2.6021, 2.6021, 2.6021,
        2.6021, 2.6021, 2.3399, 2.1104, 1.9040, 1.7141, 1.5362, 1.3670, 1.2041,
        1.0460, 0.8914, 0.7394, 0.5893, 0.4406, 0.2930, 0.1462]) return=
68694.09895647192
probs of actions:  tensor([0.9995, 0.9996, 0.9996, 0.9994, 0.9993, 0.9996,
```

```
0.9994, 0.9992, 0.9994,
        0.9991, 0.9995, 0.9990, 0.9998, 0.9997, 0.9995, 0.9998, 0.9997, 0.9993,
        0.9998, 0.9999, 0.9999, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3345, 0.3083, 0.2892, 0.2753,
0.2651, 0.2576,
        0.2520, 0.2479, 0.2448, 0.2425, 0.2407, 0.2394, 0.2385, 0.2378, 0.2372,
        0.2368, 0.2365, 0.2363, 0.2361, 0.2360, 0.2359, 0.2502])
finalReturns:  tensor([0.9944, 1.0088, 0.9935, 0.9575, 0.9066, 0.8451, 0.7758,
0.7009, 0.6218,
        0.5396, 0.4551, 0.3689, 0.2815, 0.1931, 0.1040])
----------------------------------------
iter  0  stage  9  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0047, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.7625,
2.7625, 2.7625, 2.7625, 2.7625, 2.7625, 2.7625, 2.7625, 2.7625,
        2.7625, 2.4961, 2.2637, 2.0552, 1.8638, 1.6848, 1.5149, 1.3514, 1.1929,
        1.0380, 0.8857, 0.7354, 0.5866, 0.4389, 0.2920, 0.1458]) return=
68694.09895647192
probs of actions:  tensor([0.9995, 0.9996, 0.9996, 0.9994, 0.9993, 0.9996,
0.9994, 0.9992, 0.9994,
        0.9991, 0.9995, 0.9990, 0.9998, 0.9997, 0.9995, 0.9998, 0.9997, 0.9993,
        0.9998, 0.9999, 0.9999, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3345, 0.3083, 0.2892, 0.2753,
0.2651, 0.2576,
        0.2520, 0.2479, 0.2448, 0.2425, 0.2407, 0.2394, 0.2385, 0.2378, 0.2372,
        0.2368, 0.2365, 0.2363, 0.2361, 0.2360, 0.2359, 0.2502])
finalReturns:  tensor([1.0860, 1.1004, 1.0850, 1.0487, 0.9976, 0.9359, 0.8664,
0.7913, 0.7121,
        0.6298, 0.5453, 0.4591, 0.3716, 0.2832, 0.1941, 0.1044])
----------------------------------------
iter  0  stage  8  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0057, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.9276,
2.9276, 2.9276, 2.9276, 2.9276, 2.9276, 2.9276, 2.9276, 2.9276,
        2.6556, 2.4193, 2.2080, 2.0147, 1.8342, 1.6632, 1.4990, 1.3399, 1.1845,
        1.0319, 0.8814, 0.7324, 0.5846, 0.4376, 0.2913, 0.1455]) return=
68694.09895647192
probs of actions:  tensor([0.9995, 0.9996, 0.9996, 0.9994, 0.9993, 0.9996,
0.9994, 0.9992, 0.9994,
        0.9991, 0.9995, 0.9990, 0.9998, 0.9997, 0.9995, 0.9998, 0.9997, 0.9993,
        0.9998, 0.9999, 0.9999, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
```

```
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3345, 0.3083, 0.2892, 0.2753,
0.2651, 0.2576,
        0.2520, 0.2479, 0.2448, 0.2425, 0.2407, 0.2394, 0.2385, 0.2378, 0.2372,
        0.2368, 0.2365, 0.2363, 0.2361, 0.2360, 0.2359, 0.2502])
finalReturns:  tensor([1.1785, 1.1929, 1.1773, 1.1406, 1.0892, 1.0272, 0.9575,
0.8823, 0.8029,
        0.7205, 0.6359, 0.5496, 0.4620, 0.3736, 0.2844, 0.1948, 0.1047])
----------------------------------------
iter  0  stage  7  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0070, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.0992,
3.0992, 3.0992, 3.0992, 3.0992, 3.0992, 3.0992, 3.0992, 2.8197,
        2.5781, 2.3632, 2.1671, 1.9848, 1.8123, 1.6471, 1.4872, 1.3313, 1.1783,
        1.0274, 0.8782, 0.7302, 0.5831, 0.4367, 0.2908, 0.1453]) return=
68694.09895647192
probs of actions:  tensor([0.9995, 0.9996, 0.9996, 0.9994, 0.9993, 0.9996,
0.9994, 0.9992, 0.9994,
        0.9991, 0.9995, 0.9990, 0.9998, 0.9997, 0.9995, 0.9998, 0.9997, 0.9993,
        0.9998, 0.9999, 0.9999, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3345, 0.3083, 0.2892, 0.2753,
0.2651, 0.2576,
        0.2520, 0.2479, 0.2448, 0.2425, 0.2407, 0.2394, 0.2385, 0.2378, 0.2372,
        0.2368, 0.2365, 0.2363, 0.2361, 0.2360, 0.2359, 0.2502])
finalReturns:  tensor([1.2720, 1.2864, 1.2704, 1.2334, 1.1815, 1.1191, 1.0491,
0.9736, 0.8941,
        0.8115, 0.7268, 0.6404, 0.5528, 0.4643, 0.3751, 0.2854, 0.1953, 0.1050])
----------------------------------------
iter  0  stage  6  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0080, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.2799,
3.2799, 3.2799, 3.2799, 3.2799, 3.2799, 3.2799, 2.9902, 2.7415,
        2.5215, 2.3219, 2.1370, 1.9626, 1.7960, 1.6351, 1.4784, 1.3248, 1.1736,
        1.0241, 0.8758, 0.7286, 0.5820, 0.4360, 0.2904, 0.1451]) return=
68694.09895647192
probs of actions:  tensor([0.9995, 0.9996, 0.9996, 0.9994, 0.9993, 0.9996,
0.9994, 0.9992, 0.9994,
        0.9991, 0.9995, 0.9990, 0.9998, 0.9997, 0.9995, 0.9998, 0.9997, 0.9993,
        0.9998, 0.9999, 0.9999, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3345, 0.3083, 0.2892, 0.2753,
0.2651, 0.2576,
        0.2520, 0.2479, 0.2448, 0.2425, 0.2407, 0.2394, 0.2385, 0.2378, 0.2372,
        0.2368, 0.2365, 0.2363, 0.2361, 0.2360, 0.2359, 0.2502])
```

```
finalReturns:  tensor([1.3666, 1.3810, 1.3646, 1.3270, 1.2746, 1.2117, 1.1413,
1.0654, 0.9856,
        0.9029, 0.8179, 0.7314, 0.6437, 0.5552, 0.4659, 0.3762, 0.2861, 0.1957,
        0.1051])
----------------------------------------
iter  0  stage  5  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0089, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.4731,
3.4731, 3.4731, 3.4731, 3.4731, 3.4731, 3.1695, 2.9111, 2.6844,
        2.4799, 2.2915, 2.1146, 1.9461, 1.7839, 1.6262, 1.4718, 1.3200, 1.1701,
        1.0215, 0.8740, 0.7273, 0.5812, 0.4355, 0.2901, 0.1450]) return=
68694.09895647192
probs of actions:  tensor([0.9995, 0.9996, 0.9996, 0.9994, 0.9993, 0.9996,
0.9994, 0.9992, 0.9994,
        0.9991, 0.9995, 0.9990, 0.9998, 0.9997, 0.9995, 0.9998, 0.9997, 0.9993,
        0.9998, 0.9999, 0.9999, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3345, 0.3083, 0.2892, 0.2753,
0.2651, 0.2576,
        0.2520, 0.2479, 0.2448, 0.2425, 0.2407, 0.2394, 0.2385, 0.2378, 0.2372,
        0.2368, 0.2365, 0.2363, 0.2361, 0.2360, 0.2359, 0.2502])
finalReturns:  tensor([1.4627, 1.4771, 1.4601, 1.4217, 1.3686, 1.3050, 1.2341,
1.1578, 1.0776,
        0.9945, 0.9094, 0.8228, 0.7349, 0.6463, 0.5570, 0.4672, 0.3770, 0.2866,
        0.1960, 0.1053])
----------------------------------------
iter  0  stage  4  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0102, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.6835,
3.6835, 3.6835, 3.6835, 3.6835, 3.3608, 3.0893, 2.8534, 2.6424,
        2.4492, 2.2689, 2.0980, 1.9339, 1.7748, 1.6195, 1.4669, 1.3164, 1.1675,
        1.0197, 0.8727, 0.7264, 0.5805, 0.4351, 0.2899, 0.1449]) return=
68694.09895647192
probs of actions:  tensor([0.9995, 0.9996, 0.9996, 0.9994, 0.9993, 0.9996,
0.9994, 0.9992, 0.9994,
        0.9991, 0.9995, 0.9990, 0.9998, 0.9997, 0.9995, 0.9998, 0.9997, 0.9993,
        0.9998, 0.9999, 0.9999, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3345, 0.3083, 0.2892, 0.2753,
0.2651, 0.2576,
        0.2520, 0.2479, 0.2448, 0.2425, 0.2407, 0.2394, 0.2385, 0.2378, 0.2372,
        0.2368, 0.2365, 0.2363, 0.2361, 0.2360, 0.2359, 0.2502])
finalReturns:  tensor([1.5606, 1.5750, 1.5572, 1.5179, 1.4637, 1.3993, 1.3276,
1.2507, 1.1700,
```

```
        1.0866, 1.0012, 0.9143, 0.8264, 0.7376, 0.6482, 0.5583, 0.4681, 0.3777,
        0.2870, 0.1962, 0.1054])
----------------------------------------
iter  0  stage  3  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0114, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.9177,
3.9177, 3.9177, 3.9177, 3.5687, 3.2793, 3.0307, 2.8109, 2.6114,
        2.4265, 2.2522, 2.0856, 1.9247, 1.7680, 1.6145, 1.4632, 1.3137, 1.1655,
        1.0182, 0.8717, 0.7257, 0.5801, 0.4348, 0.2897, 0.1448]) return=
68694.09895647192
probs of actions:  tensor([0.9995, 0.9996, 0.9996, 0.9994, 0.9993, 0.9996,
0.9994, 0.9992, 0.9994,
        0.9991, 0.9995, 0.9990, 0.9998, 0.9997, 0.9995, 0.9998, 0.9997, 0.9993,
        0.9998, 0.9999, 0.9999, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3345, 0.3083, 0.2892, 0.2753,
0.2651, 0.2576,
        0.2520, 0.2479, 0.2448, 0.2425, 0.2407, 0.2394, 0.2385, 0.2378, 0.2372,
        0.2368, 0.2365, 0.2363, 0.2361, 0.2360, 0.2359, 0.2502])
finalReturns:  tensor([1.6609, 1.6753, 1.6565, 1.6158, 1.5603, 1.4947, 1.4220,
1.3444, 1.2631,
        1.1792, 1.0934, 1.0062, 0.9180, 0.8291, 0.7395, 0.6496, 0.5593, 0.4688,
        0.3781, 0.2873, 0.1964, 0.1054])
----------------------------------------
iter  0  stage  2  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0123, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.1854,
4.1854, 4.1854, 3.7998, 3.4854, 3.2196, 2.9877, 2.7795, 2.5884,
        2.4096, 2.2397, 2.0764, 1.9179, 1.7630, 1.6108, 1.4605, 1.3117, 1.1640,
        1.0172, 0.8709, 0.7252, 0.5797, 0.4345, 0.2896, 0.1447]) return=
68694.09895647192
probs of actions:  tensor([0.9995, 0.9996, 0.9996, 0.9994, 0.9993, 0.9996,
0.9994, 0.9992, 0.9994,
        0.9991, 0.9995, 0.9990, 0.9998, 0.9997, 0.9995, 0.9998, 0.9997, 0.9993,
        0.9998, 0.9999, 0.9999, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3345, 0.3083, 0.2892, 0.2753,
0.2651, 0.2576,
        0.2520, 0.2479, 0.2448, 0.2425, 0.2407, 0.2394, 0.2385, 0.2378, 0.2372,
        0.2368, 0.2365, 0.2363, 0.2361, 0.2360, 0.2359, 0.2502])
finalReturns:  tensor([1.7644, 1.7788, 1.7586, 1.7161, 1.6589, 1.5917, 1.5177,
1.4390, 1.3568,
        1.2723, 1.1860, 1.0985, 1.0100, 0.9208, 0.8311, 0.7410, 0.6506, 0.5601,
        0.4693, 0.3785, 0.2875, 0.1965, 0.1055])
```

```
---------------------------------------
iter  0  stage  1  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0133, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.5005,
4.5005, 4.0633, 3.7141, 3.4244, 3.1758, 2.9558, 2.7562, 2.5713,
        2.3970, 2.2304, 2.0695, 1.9128, 1.7592, 1.6080, 1.4584, 1.3102, 1.1629,
        1.0164, 0.8704, 0.7248, 0.5795, 0.4344, 0.2895, 0.1447]) return=
68694.09895647192
probs of actions:  tensor([0.9995, 0.9996, 0.9996, 0.9994, 0.9993, 0.9996,
0.9994, 0.9992, 0.9994,
        0.9992, 0.9995, 0.9990, 0.9998, 0.9997, 0.9995, 0.9998, 0.9997, 0.9993,
        0.9998, 0.9999, 0.9999, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3345, 0.3083, 0.2892, 0.2753,
0.2651, 0.2576,
        0.2520, 0.2479, 0.2448, 0.2425, 0.2407, 0.2394, 0.2385, 0.2378, 0.2372,
        0.2368, 0.2365, 0.2363, 0.2361, 0.2360, 0.2359, 0.2502])
finalReturns:  tensor([1.8720, 1.8864, 1.8645, 1.8196, 1.7600, 1.6907, 1.6150,
1.5348, 1.4516,
        1.3662, 1.2792, 1.1911, 1.1022, 1.0128, 0.9228, 0.8326, 0.7421, 0.6514,
        0.5606, 0.4697, 0.3787, 0.2877, 0.1966, 0.1055])
---------------------------------------
iter  0  stage  0  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0145, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.8841,
4.3729, 3.9744, 3.6513, 3.3795, 3.1433, 2.9322, 2.7389, 2.5586,
        2.3875, 2.2234, 2.0643, 1.9089, 1.7564, 1.6059, 1.4569, 1.3091, 1.1621,
        1.0158, 0.8699, 0.7245, 0.5793, 0.4343, 0.2894, 0.1447]) return=
68694.09895647192
probs of actions:  tensor([0.9995, 0.9996, 0.9996, 0.9994, 0.9993, 0.9996,
0.9994, 0.9992, 0.9994,
        0.9992, 0.9995, 0.9990, 0.9998, 0.9997, 0.9995, 0.9998, 0.9997, 0.9993,
        0.9998, 0.9999, 0.9999, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3345, 0.3083, 0.2892, 0.2753,
0.2651, 0.2576,
        0.2520, 0.2479, 0.2448, 0.2425, 0.2407, 0.2394, 0.2385, 0.2378, 0.2372,
        0.2368, 0.2365, 0.2363, 0.2361, 0.2360, 0.2359, 0.2502])
finalReturns:  tensor([1.9853, 1.9997, 1.9753, 1.9272, 1.8645, 1.7924, 1.7143,
1.6323, 1.5475,
        1.4610, 1.3731, 1.2844, 1.1950, 1.1051, 1.0148, 0.9244, 0.8337, 0.7429,
        0.6520, 0.5610, 0.4700, 0.3789, 0.2878, 0.1967, 0.1056])
0,[1e-05,1][1, 10000, 1, 1],1682652460 saved
[891910, 'tensor([0., 0., 1., 0., 0., 0., 0., 0., 0., 0., 0., 0.])',
```

```
68694.09895647192, 88516.4319236399, 0.014486568048596382, 1e-05, 1, 0,
'tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12,\n        12, 12, 12, 12, 12, 12,  0])', '[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.\n 1.]', '0,[1e-05,1][1, 10000, 1,
1],1682652460', 25, 50, 157611.33342506486, 175538.2609849781, 68707.8951114289,
135313.23466666666, 132439.05866666668, 122329.26544005591, 122329.26544005591,
130254.64915679736, 130254.64915679736, 80045.82189636558, 122329.26544005591,
130254.64915679736]
policy reset
----------------------------------------
iter  1  stage  24  ep  99999   adversary:  AdversaryModes.constant_95
  actions:  tensor([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0,
        0])
loss=  tensor(-0., grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.1455,
0.1455, 0.1455, 0.1455, 0.1455, 0.1455, 0.1455, 0.1455, 0.1455,
        0.1455, 0.1455, 0.1455, 0.1455, 0.1455, 0.1455, 0.1455, 0.1455, 0.1455,
        0.1455, 0.1455, 0.1455, 0.1455, 0.1455, 0.1455, 0.1455]) return=
48892.64112506197
probs of actions:  tensor([0.8787, 0.8940, 0.8694, 0.8601, 0.8593, 0.8601,
0.8369, 0.8052, 0.8556,
        0.8616, 0.8490, 0.8605, 0.8540, 0.8323, 0.8056, 0.8109, 0.8350, 0.8404,
        0.8529, 0.8607, 0.0669, 0.8381, 0.8490, 0.8036, 0.9799],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5112, 0.3985, 0.3231, 0.2718, 0.2362, 0.2111, 0.1933,
0.1804, 0.1710,
        0.1641, 0.1591, 0.1554, 0.1526, 0.1505, 0.1490, 0.1478, 0.1470, 0.1463,
        0.1458, 0.1455, 0.1451, 0.1469, 0.1463, 0.1458, 0.1455])
finalReturns:  tensor([0.])
----------------------------------------
iter  1  stage  23  ep  99999   adversary:  AdversaryModes.constant_95
  actions:  tensor([ 5,  0,  9,  6,  4,  4,  3,  6,  9,  6,  7,  6,  9,  6, 15,
9,  6,  6,
         5,  5,  6,  7,  5,  9,  0])
loss=  tensor(0.0071, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.3767,
0.3767, 0.3767, 0.3767, 0.3767, 0.3767, 0.3767, 0.3767, 0.3767,
        0.3767, 0.3767, 0.3767, 0.3767, 0.3767, 0.3767, 0.3767, 0.3767, 0.3767,
        0.3767, 0.3767, 0.3767, 0.3767, 0.3767, 0.3767, 0.1817]) return=
59134.4246096995
probs of actions:  tensor([0.1258, 0.1556, 0.3641, 0.1753, 0.0548, 0.0523,
0.0173, 0.1730, 0.3084,
        0.1926, 0.1241, 0.1785, 0.2978, 0.1695, 0.0034, 0.2990, 0.1839, 0.1763,
        0.1376, 0.1150, 0.2031, 0.1178, 0.1148, 0.5464, 0.9951],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5087, 0.4144, 0.3258, 0.2998, 0.2721, 0.2458, 0.2276,
0.2089, 0.1996,
        0.2074, 0.2017, 0.2019, 0.1944, 0.2034, 0.1811, 0.2137, 0.2180, 0.2108,
        0.2066, 0.2004, 0.1947, 0.1922, 0.1960, 0.1869, 0.2014])
```

```
finalReturns:  tensor([0.0116, 0.0197])
----------------------------------------
iter  1  stage  22  ep  99999   adversary:  AdversaryModes.constant_95
  actions:  tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 7, 9, 9,
9, 9, 9, 9,
        0])
loss=  tensor(0.0002, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.6042,
0.6042, 0.6042, 0.6042, 0.6042, 0.6042, 0.6042, 0.6042, 0.6042, 0.6042,
        0.6042, 0.6042, 0.6042, 0.6042, 0.6042, 0.6042, 0.6042, 0.6042, 0.6042,
        0.6042, 0.6042, 0.6042, 0.6042, 0.6042, 0.6042, 0.3844, 0.1849]) return=
63727.41292382953
probs of actions:  tensor([0.9949, 0.9901, 0.9940, 0.9882, 0.9920, 0.9925,
0.9934, 0.9894, 0.9914,
        0.9911, 0.9930, 0.9908, 0.9882, 0.9847, 0.9907, 0.9878, 0.9905, 0.0035,
        0.9919, 0.9858, 0.9927, 0.9911, 0.9983, 0.9988, 0.9993],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2365,
        0.2304, 0.2260, 0.2226, 0.2202, 0.2183, 0.2169, 0.2159, 0.2151, 0.2177,
        0.2094, 0.2103, 0.2109, 0.2114, 0.2117, 0.2120, 0.2203])
finalReturns:  tensor([0.0398, 0.0479, 0.0354])
----------------------------------------
iter  1  stage  22  ep  113089   adversary:  AdversaryModes.constant_95
  actions:  tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 0, 9, 9,
        0])
loss=  tensor(7.9259e-05, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([0.5613, 0.5613, 0.5613, 0.5613, 0.5613, 0.5613, 0.5613, 0.5613, 0.5613,
        0.5613, 0.5613, 0.5613, 0.5613, 0.5613, 0.5613, 0.5613, 0.5613, 0.5613,
        0.5613, 0.5613, 0.5613, 0.5613, 0.5613, 0.3606, 0.1749]) return=
63459.11549104476
probs of actions:  tensor([9.9744e-01, 9.9532e-01, 9.9700e-01, 9.9432e-01,
9.9599e-01, 9.9599e-01,
        9.9645e-01, 9.9433e-01, 9.9540e-01, 9.9529e-01, 9.9630e-01, 9.9524e-01,
        9.9393e-01, 9.9210e-01, 9.9491e-01, 9.9326e-01, 9.9492e-01, 9.9538e-01,
        9.9567e-01, 9.9287e-01, 9.9599e-01, 8.9096e-04, 9.9901e-01, 9.9946e-01,
        9.9952e-01], grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2365,
        0.2304, 0.2260, 0.2226, 0.2202, 0.2183, 0.2169, 0.2159, 0.2151, 0.2145,
        0.2141, 0.2138, 0.2135, 0.2214, 0.1925, 0.1975, 0.2094])
finalReturns:  tensor([0.0382, 0.0463, 0.0345])
----------------------------------------
iter  1  stage  21  ep  25407   adversary:  AdversaryModes.constant_95
  actions:  tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9,
        0])
loss=  tensor(0.0001, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.7827,
```

```
0.7827, 0.7827, 0.7827, 0.7827, 0.7827, 0.7827, 0.7827, 0.7827,
        0.7827, 0.7827, 0.7827, 0.7827, 0.7827, 0.7827, 0.7827, 0.7827, 0.7827,
        0.7827, 0.7827, 0.7827, 0.7827, 0.5613, 0.3606, 0.1749]) return=
63858.06679267008
probs of actions:  tensor([0.9994, 0.9990, 0.9993, 0.9987, 0.9991, 0.9989,
0.9990, 0.9985, 0.9987,
        0.9987, 0.9990, 0.9988, 0.9985, 0.9981, 0.9986, 0.9981, 0.9986, 0.9988,
        0.9988, 0.9983, 0.9989, 0.9990, 0.9998, 1.0000, 0.9996],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2365,
        0.2304, 0.2260, 0.2226, 0.2202, 0.2183, 0.2169, 0.2159, 0.2151, 0.2145,
        0.2141, 0.2138, 0.2135, 0.2133, 0.2132, 0.2131, 0.2211])
finalReturns:  tensor([0.0781, 0.0862, 0.0736, 0.0462])
----------------------------------------
iter  1  stage  20  ep  42   adversary:  AdversaryModes.constant_95
  actions:  tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9,
        0])
loss=  tensor(0.0003, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.9502,
0.9502, 0.9502, 0.9502, 0.9502, 0.9502, 0.9502, 0.9502, 0.9502,
        0.9502, 0.9502, 0.9502, 0.9502, 0.9502, 0.9502, 0.9502, 0.9502, 0.9502,
        0.9502, 0.9502, 0.9502, 0.7286, 0.5278, 0.3420, 0.1671]) return=
63858.06679267008
probs of actions:  tensor([0.9994, 0.9990, 0.9993, 0.9987, 0.9991, 0.9989,
0.9991, 0.9985, 0.9987,
        0.9987, 0.9990, 0.9988, 0.9985, 0.9981, 0.9986, 0.9982, 0.9987, 0.9988,
        0.9988, 0.9983, 0.9990, 0.9990, 0.9998, 1.0000, 0.9996],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2365,
        0.2304, 0.2260, 0.2226, 0.2202, 0.2183, 0.2169, 0.2159, 0.2151, 0.2145,
        0.2141, 0.2138, 0.2135, 0.2133, 0.2132, 0.2131, 0.2211])
finalReturns:  tensor([0.1241, 0.1322, 0.1196, 0.0922, 0.0541])
----------------------------------------
iter  1  stage  19  ep  280   adversary:  AdversaryModes.constant_95
  actions:  tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9,
        0])
loss=  tensor(0.0005, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.1122,
1.1122, 1.1122, 1.1122, 1.1122, 1.1122, 1.1122, 1.1122, 1.1122,
        1.1122, 1.1122, 1.1122, 1.1122, 1.1122, 1.1122, 1.1122, 1.1122, 1.1122,
        1.1122, 1.1122, 0.8904, 0.6894, 0.5035, 0.3284, 0.1613]) return=
63858.06679267008
probs of actions:  tensor([0.9995, 0.9992, 0.9994, 0.9989, 0.9992, 0.9991,
0.9992, 0.9987, 0.9989,
        0.9989, 0.9992, 0.9990, 0.9987, 0.9985, 0.9989, 0.9985, 0.9989, 0.9990,
        0.9990, 0.9990, 0.9994, 0.9992, 0.9998, 1.0000, 0.9995],
```

```
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2365,
        0.2304, 0.2260, 0.2226, 0.2202, 0.2183, 0.2169, 0.2159, 0.2151, 0.2145,
        0.2141, 0.2138, 0.2135, 0.2133, 0.2132, 0.2131, 0.2211])
finalReturns:  tensor([0.1759, 0.1840, 0.1714, 0.1440, 0.1058, 0.0598])
----------------------------------------
iter  1  stage  18  ep  1   adversary:  AdversaryModes.constant_95
  actions:  tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9,
        0])
loss=  tensor(0.0008, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.2703,
1.2703, 1.2703, 1.2703, 1.2703, 1.2703, 1.2703, 1.2703, 1.2703,
        1.2703, 1.2703, 1.2703, 1.2703, 1.2703, 1.2703, 1.2703, 1.2703, 1.2703,
        1.2703, 1.0481, 0.8469, 0.6608, 0.4856, 0.3184, 0.1570]) return=
63858.06679267008
probs of actions:  tensor([0.9995, 0.9992, 0.9994, 0.9989, 0.9992, 0.9991,
0.9992, 0.9987, 0.9989,
        0.9990, 0.9992, 0.9990, 0.9987, 0.9985, 0.9989, 0.9985, 0.9989, 0.9990,
        0.9990, 0.9990, 0.9994, 0.9992, 0.9998, 1.0000, 0.9995],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2365,
        0.2304, 0.2260, 0.2226, 0.2202, 0.2183, 0.2169, 0.2159, 0.2151, 0.2145,
        0.2141, 0.2138, 0.2135, 0.2133, 0.2132, 0.2131, 0.2211])
finalReturns:  tensor([0.2319, 0.2400, 0.2274, 0.2000, 0.1619, 0.1159, 0.0641])
----------------------------------------
iter  1  stage  17  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9,
        0])
loss=  tensor(0.0013, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.4255,
1.4255, 1.4255, 1.4255, 1.4255, 1.4255, 1.4255, 1.4255, 1.4255,
        1.4255, 1.4255, 1.4255, 1.4255, 1.4255, 1.4255, 1.4255, 1.4255, 1.4255,
        1.2029, 1.0014, 0.8151, 0.6397, 0.4724, 0.3110, 0.1539]) return=
63858.06679267008
probs of actions:  tensor([0.9995, 0.9992, 0.9994, 0.9989, 0.9992, 0.9991,
0.9992, 0.9987, 0.9990,
        0.9990, 0.9992, 0.9990, 0.9987, 0.9985, 0.9989, 0.9985, 0.9989, 0.9990,
        0.9990, 0.9990, 0.9994, 0.9992, 0.9998, 1.0000, 0.9995],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2365,
        0.2304, 0.2260, 0.2226, 0.2202, 0.2183, 0.2169, 0.2159, 0.2151, 0.2145,
        0.2141, 0.2138, 0.2135, 0.2133, 0.2132, 0.2131, 0.2211])
finalReturns:  tensor([0.2912, 0.2993, 0.2867, 0.2593, 0.2211, 0.1751, 0.1233,
0.0673])
----------------------------------------
```

```
iter  1  stage  16  ep  27   adversary:  AdversaryModes.constant_95
  actions:  tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9,
        0])
loss=  tensor(0.0017, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.5790,
1.5790, 1.5790, 1.5790, 1.5790, 1.5790, 1.5790, 1.5790, 1.5790,
        1.5790, 1.5790, 1.5790, 1.5790, 1.5790, 1.5790, 1.5790, 1.5790, 1.3558,
        1.1539, 0.9672, 0.7917, 0.6242, 0.4626, 0.3055, 0.1515]) return=
63858.06679267008
probs of actions:  tensor([0.9995, 0.9992, 0.9995, 0.9990, 0.9993, 0.9992,
0.9993, 0.9988, 0.9990,
        0.9990, 0.9992, 0.9991, 0.9988, 0.9985, 0.9989, 0.9985, 0.9990, 0.9991,
        0.9991, 0.9991, 0.9994, 0.9992, 0.9998, 1.0000, 0.9995],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2365,
        0.2304, 0.2260, 0.2226, 0.2202, 0.2183, 0.2169, 0.2159, 0.2151, 0.2145,
        0.2141, 0.2138, 0.2135, 0.2133, 0.2132, 0.2131, 0.2211])
finalReturns:  tensor([0.3529, 0.3610, 0.3483, 0.3209, 0.2827, 0.2366, 0.1848,
0.1288, 0.0696])
----------------------------------------
iter  1  stage  15  ep  96   adversary:  AdversaryModes.constant_95
  actions:  tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9,
        0])
loss=  tensor(0.0018, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.7314,
1.7314, 1.7314, 1.7314, 1.7314, 1.7314, 1.7314, 1.7314, 1.7314,
        1.7314, 1.7314, 1.7314, 1.7314, 1.7314, 1.7314, 1.7314, 1.5074, 1.3050,
        1.1179, 0.9421, 0.7744, 0.6127, 0.4554, 0.3014, 0.1498]) return=
63858.06679267008
probs of actions:  tensor([0.9996, 0.9994, 0.9996, 0.9992, 0.9994, 0.9993,
0.9994, 0.9990, 0.9992,
        0.9992, 0.9994, 0.9993, 0.9991, 0.9988, 0.9991, 0.9990, 0.9993, 0.9994,
        0.9994, 0.9993, 0.9996, 0.9994, 0.9999, 1.0000, 0.9995],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2365,
        0.2304, 0.2260, 0.2226, 0.2202, 0.2183, 0.2169, 0.2159, 0.2151, 0.2145,
        0.2141, 0.2138, 0.2135, 0.2133, 0.2132, 0.2131, 0.2211])
finalReturns:  tensor([0.4163, 0.4244, 0.4118, 0.3843, 0.3460, 0.3000, 0.2481,
0.1921, 0.1329,
        0.0714])
----------------------------------------
iter  1  stage  14  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9,
        0])
loss=  tensor(0.0024, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.8835,
```

```
1.8835, 1.8835, 1.8835, 1.8835, 1.8835, 1.8835, 1.8835, 1.8835,
        1.8835, 1.8835, 1.8835, 1.8835, 1.8835, 1.8835, 1.6585, 1.4553, 1.2677,
        1.0915, 0.9235, 0.7615, 0.6041, 0.4500, 0.2983, 0.1485]) return=
63858.06679267008
probs of actions:  tensor([0.9996, 0.9994, 0.9996, 0.9992, 0.9994, 0.9993,
0.9994, 0.9990, 0.9992,
        0.9992, 0.9994, 0.9993, 0.9991, 0.9988, 0.9991, 0.9990, 0.9993, 0.9994,
        0.9994, 0.9993, 0.9996, 0.9994, 0.9999, 1.0000, 0.9994],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2365,
        0.2304, 0.2260, 0.2226, 0.2202, 0.2183, 0.2169, 0.2159, 0.2151, 0.2145,
        0.2141, 0.2138, 0.2135, 0.2133, 0.2132, 0.2131, 0.2211])
finalReturns:  tensor([0.4812, 0.4893, 0.4766, 0.4490, 0.4108, 0.3646, 0.3128,
0.2567, 0.1975,
        0.1359, 0.0727])
----------------------------------------
iter  1  stage  13  ep  29   adversary:  AdversaryModes.constant_95
  actions:  tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9,
        0])
loss=  tensor(0.0028, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.0358,
2.0358, 2.0358, 2.0358, 2.0358, 2.0358, 2.0358, 2.0358, 2.0358,
        2.0358, 2.0358, 2.0358, 2.0358, 2.0358, 1.8094, 1.6052, 1.4170, 1.2402,
        1.0719, 0.9097, 0.7520, 0.5977, 0.4459, 0.2960, 0.1475]) return=
63858.06679267008
probs of actions:  tensor([0.9997, 0.9994, 0.9996, 0.9993, 0.9995, 0.9994,
0.9995, 0.9991, 0.9993,
        0.9993, 0.9995, 0.9993, 0.9991, 0.9990, 0.9993, 0.9991, 0.9994, 0.9995,
        0.9994, 0.9994, 0.9996, 0.9994, 0.9999, 1.0000, 0.9994],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2365,
        0.2304, 0.2260, 0.2226, 0.2202, 0.2183, 0.2169, 0.2159, 0.2151, 0.2145,
        0.2141, 0.2138, 0.2135, 0.2133, 0.2132, 0.2131, 0.2211])
finalReturns:  tensor([0.5471, 0.5552, 0.5425, 0.5149, 0.4765, 0.4303, 0.3784,
0.3223, 0.2631,
        0.2015, 0.1382, 0.0736])
----------------------------------------
iter  1  stage  12  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9,
        0])
loss=  tensor(0.0035, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.1892,
2.1892, 2.1892, 2.1892, 2.1892, 2.1892, 2.1892, 2.1892, 2.1892,
        2.1892, 2.1892, 2.1892, 2.1892, 1.9609, 1.7554, 1.5662, 1.3887, 1.2199,
        1.0573, 0.8994, 0.7449, 0.5930, 0.4429, 0.2943, 0.1468]) return=
63858.06679267008
```

```
probs of actions:  tensor([0.9997, 0.9994, 0.9996, 0.9993, 0.9995, 0.9994,
0.9995, 0.9991, 0.9993,
        0.9993, 0.9995, 0.9993, 0.9991, 0.9990, 0.9993, 0.9991, 0.9994, 0.9995,
        0.9994, 0.9994, 0.9996, 0.9994, 0.9999, 1.0000, 0.9994],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2365,
        0.2304, 0.2260, 0.2226, 0.2202, 0.2183, 0.2169, 0.2159, 0.2151, 0.2145,
        0.2141, 0.2138, 0.2135, 0.2133, 0.2132, 0.2131, 0.2211])
finalReturns:  tensor([0.6140, 0.6221, 0.6093, 0.5815, 0.5431, 0.4969, 0.4449,
0.3887, 0.3294,
        0.2678, 0.2045, 0.1399, 0.0744])
----------------------------------------
iter  1  stage  11  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9,
        0])
loss=  tensor(0.0042, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.3442,
2.3442, 2.3442, 2.3442, 2.3442, 2.3442, 2.3442, 2.3442, 2.3442,
        2.3442, 2.3442, 2.3442, 2.1135, 1.9062, 1.7157, 1.5374, 1.3678, 1.2048,
        1.0465, 0.8917, 0.7396, 0.5894, 0.4407, 0.2930, 0.1462]) return=
63858.06679267008
probs of actions:  tensor([0.9997, 0.9994, 0.9996, 0.9993, 0.9995, 0.9994,
0.9995, 0.9991, 0.9993,
        0.9993, 0.9995, 0.9993, 0.9991, 0.9990, 0.9993, 0.9991, 0.9994, 0.9995,
        0.9994, 0.9994, 0.9996, 0.9994, 0.9999, 1.0000, 0.9994],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2365,
        0.2304, 0.2260, 0.2226, 0.2202, 0.2183, 0.2169, 0.2159, 0.2151, 0.2145,
        0.2141, 0.2138, 0.2135, 0.2133, 0.2132, 0.2131, 0.2211])
finalReturns:  tensor([0.6815, 0.6896, 0.6767, 0.6489, 0.6104, 0.5640, 0.5120,
0.4557, 0.3964,
        0.3347, 0.2714, 0.2068, 0.1412, 0.0749])
----------------------------------------
iter  1  stage  10  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9,
        0])
loss=  tensor(0.0048, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.5019,
2.5019, 2.5019, 2.5019, 2.5019, 2.5019, 2.5019, 2.5019, 2.5019,
        2.5019, 2.5019, 2.2679, 2.0583, 1.8661, 1.6865, 1.5160, 1.3523, 1.1935,
        1.0384, 0.8860, 0.7356, 0.5867, 0.4390, 0.2921, 0.1458]) return=
63858.06679267008
probs of actions:  tensor([0.9997, 0.9994, 0.9996, 0.9993, 0.9995, 0.9994,
0.9995, 0.9991, 0.9993,
        0.9993, 0.9995, 0.9994, 0.9991, 0.9990, 0.9993, 0.9991, 0.9994, 0.9995,
        0.9994, 0.9994, 0.9996, 0.9994, 0.9999, 1.0000, 0.9994],
```

```
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2365,
        0.2304, 0.2260, 0.2226, 0.2202, 0.2183, 0.2169, 0.2159, 0.2151, 0.2145,
        0.2141, 0.2138, 0.2135, 0.2133, 0.2132, 0.2131, 0.2211])
finalReturns:  tensor([0.7498, 0.7579, 0.7449, 0.7169, 0.6782, 0.6317, 0.5795,
0.5232, 0.4638,
        0.4021, 0.3387, 0.2740, 0.2085, 0.1422, 0.0753])
----------------------------------------
iter  1  stage  9  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9,
        0])
loss=  tensor(0.0055, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.6635,
2.6635, 2.6635, 2.6635, 2.6635, 2.6635, 2.6635, 2.6635, 2.6635,
        2.6635, 2.4249, 2.2121, 2.0177, 1.8364, 1.6648, 1.5002, 1.3408, 1.1852,
        1.0324, 0.8817, 0.7327, 0.5848, 0.4377, 0.2914, 0.1455]) return=
63858.06679267008
probs of actions:  tensor([0.9997, 0.9994, 0.9996, 0.9993, 0.9995, 0.9994,
0.9995, 0.9991, 0.9993,
        0.9993, 0.9995, 0.9994, 0.9991, 0.9990, 0.9993, 0.9991, 0.9994, 0.9995,
        0.9994, 0.9994, 0.9996, 0.9994, 0.9999, 1.0000, 0.9994],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2365,
        0.2304, 0.2260, 0.2226, 0.2202, 0.2183, 0.2169, 0.2159, 0.2151, 0.2145,
        0.2141, 0.2138, 0.2135, 0.2133, 0.2132, 0.2131, 0.2211])
finalReturns:  tensor([0.8187, 0.8268, 0.8136, 0.7854, 0.7465, 0.6999, 0.6475,
0.5911, 0.5316,
        0.4698, 0.4064, 0.3417, 0.2760, 0.2097, 0.1429, 0.0756])
----------------------------------------
iter  1  stage  8  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9,
        0])
loss=  tensor(0.0063, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.8303,
2.8303, 2.8303, 2.8303, 2.8303, 2.8303, 2.8303, 2.8303, 2.8303,
        2.5858, 2.3687, 2.1712, 1.9877, 1.8145, 1.6487, 1.4884, 1.3321, 1.1789,
        1.0279, 0.8785, 0.7304, 0.5833, 0.4368, 0.2908, 0.1453]) return=
63858.06679267008
probs of actions:  tensor([0.9997, 0.9994, 0.9996, 0.9993, 0.9995, 0.9994,
0.9995, 0.9991, 0.9993,
        0.9993, 0.9995, 0.9994, 0.9992, 0.9990, 0.9993, 0.9991, 0.9994, 0.9995,
        0.9994, 0.9994, 0.9996, 0.9994, 0.9999, 1.0000, 0.9994],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2365,
        0.2304, 0.2260, 0.2226, 0.2202, 0.2183, 0.2169, 0.2159, 0.2151, 0.2145,
```

```
        0.2141, 0.2138, 0.2135, 0.2133, 0.2132, 0.2131, 0.2211])
finalReturns:  tensor([0.8883, 0.8964, 0.8830, 0.8546, 0.8154, 0.7685, 0.7160,
0.6594, 0.5997,
        0.5378, 0.4743, 0.4096, 0.3439, 0.2775, 0.2107, 0.1434, 0.0759])
---------------------------------------
iter  1  stage  7  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9,
        0])
loss=  tensor(0.0073, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.0045,
3.0045, 3.0045, 3.0045, 3.0045, 3.0045, 3.0045, 3.0045, 2.7518,
        2.5290, 2.3274, 2.1410, 1.9656, 1.7982, 1.6367, 1.4796, 1.3257, 1.1742,
        1.0245, 0.8761, 0.7288, 0.5822, 0.4361, 0.2904, 0.1451]) return=
63858.06679267008
probs of actions:  tensor([0.9997, 0.9995, 0.9996, 0.9993, 0.9995, 0.9994,
0.9995, 0.9991, 0.9993,
        0.9993, 0.9995, 0.9994, 0.9992, 0.9990, 0.9993, 0.9991, 0.9994, 0.9995,
        0.9994, 0.9994, 0.9996, 0.9994, 0.9999, 1.0000, 0.9994],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2365,
        0.2304, 0.2260, 0.2226, 0.2202, 0.2183, 0.2169, 0.2159, 0.2151, 0.2145,
        0.2141, 0.2138, 0.2135, 0.2133, 0.2132, 0.2131, 0.2211])
finalReturns:  tensor([0.9587, 0.9668, 0.9531, 0.9243, 0.8848, 0.8376, 0.7848,
0.7280, 0.6682,
        0.6062, 0.5425, 0.4777, 0.4120, 0.3456, 0.2786, 0.2114, 0.1438, 0.0760])
---------------------------------------
iter  1  stage  6  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9,
        0])
loss=  tensor(0.0080, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.1888,
3.1888, 3.1888, 3.1888, 3.1888, 3.1888, 3.1888, 2.9251, 2.6945,
        2.4873, 2.2969, 2.1186, 1.9491, 1.7860, 1.6278, 1.4730, 1.3209, 1.1707,
        1.0220, 0.8744, 0.7275, 0.5813, 0.4356, 0.2901, 0.1450]) return=
63858.06679267008
probs of actions:  tensor([0.9997, 0.9995, 0.9996, 0.9993, 0.9995, 0.9994,
0.9995, 0.9991, 0.9993,
        0.9993, 0.9995, 0.9994, 0.9992, 0.9990, 0.9993, 0.9991, 0.9994, 0.9995,
        0.9994, 0.9994, 0.9996, 0.9994, 0.9999, 1.0000, 0.9994],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2365,
        0.2304, 0.2260, 0.2226, 0.2202, 0.2183, 0.2169, 0.2159, 0.2151, 0.2145,
        0.2141, 0.2138, 0.2135, 0.2133, 0.2132, 0.2131, 0.2211])
finalReturns:  tensor([1.0301, 1.0382, 1.0241, 0.9949, 0.9548, 0.9072, 0.8541,
0.7969, 0.7369,
        0.6747, 0.6110, 0.5460, 0.4802, 0.4138, 0.3468, 0.2795, 0.2119, 0.1441,
```

```
        0.0762])
---------------------------------------
iter  1  stage  5  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9,
        0])
loss=  tensor(0.0089, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.3871,
3.3871, 3.3871, 3.3871, 3.3871, 3.3871, 3.1081, 2.8670, 2.6523,
        2.4565, 2.2743, 2.1019, 1.9368, 1.7770, 1.6211, 1.4681, 1.3173, 1.1681,
        1.0201, 0.8730, 0.7266, 0.5807, 0.4352, 0.2899, 0.1449]) return=
63858.06679267008
probs of actions:  tensor([0.9997, 0.9995, 0.9996, 0.9993, 0.9995, 0.9994,
0.9995, 0.9991, 0.9993,
        0.9993, 0.9995, 0.9994, 0.9992, 0.9990, 0.9993, 0.9991, 0.9994, 0.9995,
        0.9995, 0.9994, 0.9996, 0.9994, 0.9999, 1.0000, 0.9994],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2365,
        0.2304, 0.2260, 0.2226, 0.2202, 0.2183, 0.2169, 0.2159, 0.2151, 0.2145,
        0.2141, 0.2138, 0.2135, 0.2133, 0.2132, 0.2131, 0.2211])
finalReturns:  tensor([1.1027, 1.1108, 1.0963, 1.0663, 1.0257, 0.9775, 0.9238,
0.8663, 0.8060,
        0.7436, 0.6797, 0.6146, 0.5487, 0.4821, 0.4151, 0.3477, 0.2801, 0.2123,
        0.1443, 0.0763])
---------------------------------------
iter  1  stage  4  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9,
        0])
loss=  tensor(0.0096, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.6046,
3.6046, 3.6046, 3.6046, 3.6046, 3.3048, 3.0491, 2.8242, 2.6211,
        2.4337, 2.2575, 2.0895, 1.9276, 1.7702, 1.6161, 1.4644, 1.3146, 1.1661,
        1.0187, 0.8720, 0.7259, 0.5802, 0.4349, 0.2897, 0.1448]) return=
63858.06679267008
probs of actions:  tensor([0.9997, 0.9995, 0.9996, 0.9993, 0.9995, 0.9994,
0.9995, 0.9991, 0.9993,
        0.9993, 0.9995, 0.9994, 0.9992, 0.9991, 0.9993, 0.9992, 0.9994, 0.9995,
        0.9995, 0.9994, 0.9996, 0.9994, 0.9999, 1.0000, 0.9994],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2365,
        0.2304, 0.2260, 0.2226, 0.2202, 0.2183, 0.2169, 0.2159, 0.2151, 0.2145,
        0.2141, 0.2138, 0.2135, 0.2133, 0.2132, 0.2131, 0.2211])
finalReturns:  tensor([1.1768, 1.1849, 1.1698, 1.1390, 1.0975, 1.0485, 0.9943,
0.9362, 0.8755,
        0.8128, 0.7486, 0.6833, 0.6173, 0.5506, 0.4835, 0.4161, 0.3484, 0.2806,
        0.2126, 0.1445, 0.0763])
---------------------------------------
```

```
iter  1  stage  3  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9,
        0])
loss=  tensor(0.0105, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.8490,
3.8490, 3.8490, 3.8490, 3.5203, 3.2446, 3.0057, 2.7927, 2.5981,
        2.4167, 2.2449, 2.0803, 1.9208, 1.7651, 1.6123, 1.4617, 1.3126, 1.1647,
        1.0176, 0.8713, 0.7254, 0.5799, 0.4346, 0.2896, 0.1447]) return=
63858.06679267008
probs of actions:  tensor([0.9997, 0.9995, 0.9996, 0.9993, 0.9995, 0.9994,
0.9995, 0.9992, 0.9993,
        0.9993, 0.9995, 0.9994, 0.9992, 0.9991, 0.9993, 0.9992, 0.9994, 0.9995,
        0.9995, 0.9994, 0.9996, 0.9995, 0.9999, 1.0000, 0.9994],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2365,
        0.2304, 0.2260, 0.2226, 0.2202, 0.2183, 0.2169, 0.2159, 0.2151, 0.2145,
        0.2141, 0.2138, 0.2135, 0.2133, 0.2132, 0.2131, 0.2211])
finalReturns:  tensor([1.2530, 1.2611, 1.2451, 1.2132, 1.1705, 1.1206, 1.0655,
1.0068, 0.9455,
        0.8824, 0.8178, 0.7523, 0.6861, 0.6193, 0.5521, 0.4846, 0.4169, 0.3490,
        0.2809, 0.2128, 0.1446, 0.0764])
----------------------------------------
iter  1  stage  2  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9,
        0])
loss=  tensor(0.0111, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.1314,
4.1314, 4.1314, 3.7620, 3.4586, 3.2004, 2.9737, 2.7693, 2.5809,
        2.4040, 2.2356, 2.0733, 1.9156, 1.7613, 1.6095, 1.4596, 1.3111, 1.1635,
        1.0168, 0.8707, 0.7250, 0.5796, 0.4345, 0.2895, 0.1447]) return=
63858.06679267008
probs of actions:  tensor([0.9997, 0.9995, 0.9996, 0.9993, 0.9995, 0.9994,
0.9995, 0.9992, 0.9993,
        0.9993, 0.9995, 0.9994, 0.9992, 0.9991, 0.9993, 0.9992, 0.9994, 0.9995,
        0.9995, 0.9994, 0.9996, 0.9995, 0.9999, 1.0000, 0.9994],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2365,
        0.2304, 0.2260, 0.2226, 0.2202, 0.2183, 0.2169, 0.2159, 0.2151, 0.2145,
        0.2141, 0.2138, 0.2135, 0.2133, 0.2132, 0.2131, 0.2211])
finalReturns:  tensor([1.3319, 1.3400, 1.3228, 1.2894, 1.2452, 1.1939, 1.1377,
1.0781, 1.0161,
        0.9524, 0.8875, 0.8216, 0.7551, 0.6881, 0.6208, 0.5532, 0.4854, 0.4174,
        0.3493, 0.2812, 0.2130, 0.1447, 0.0764])
----------------------------------------
iter  1  stage  1  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
```

```
9, 9, 9, 9,
        0])
loss=  tensor(0.0120, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.4683,
4.4683, 4.0409, 3.6983, 3.4131, 3.1676, 2.9499, 2.7519, 2.5681,
        2.3946, 2.2286, 2.0682, 1.9118, 1.7585, 1.6074, 1.4580, 1.3099, 1.1627,
        1.0162, 0.8703, 0.7247, 0.5794, 0.4343, 0.2894, 0.1447]) return=
63858.06679267008
probs of actions:  tensor([0.9997, 0.9995, 0.9996, 0.9993, 0.9995, 0.9994,
0.9995, 0.9992, 0.9993,
        0.9993, 0.9995, 0.9994, 0.9992, 0.9991, 0.9993, 0.9992, 0.9994, 0.9995,
        0.9995, 0.9994, 0.9996, 0.9995, 0.9999, 1.0000, 0.9994],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2365,
        0.2304, 0.2260, 0.2226, 0.2202, 0.2183, 0.2169, 0.2159, 0.2151, 0.2145,
        0.2141, 0.2138, 0.2135, 0.2133, 0.2132, 0.2131, 0.2211])
finalReturns:  tensor([1.4144, 1.4225, 1.4038, 1.3683, 1.3221, 1.2690, 1.2113,
1.1505, 1.0876,
        1.0231, 0.9576, 0.8913, 0.8245, 0.7572, 0.6897, 0.6219, 0.5540, 0.4860,
        0.4178, 0.3496, 0.2814, 0.2131, 0.1448, 0.0765])
----------------------------------------
iter  1  stage  0  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9,
        0])
loss=  tensor(0.0125, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.8841,
4.3729, 3.9744, 3.6513, 3.3795, 3.1433, 2.9322, 2.7389, 2.5586,
        2.3875, 2.2234, 2.0643, 1.9089, 1.7564, 1.6059, 1.4569, 1.3091, 1.1621,
        1.0158, 0.8699, 0.7245, 0.5793, 0.4343, 0.2894, 0.1447]) return=
63858.06679267008
probs of actions:  tensor([0.9997, 0.9995, 0.9996, 0.9993, 0.9995, 0.9994,
0.9995, 0.9992, 0.9993,
        0.9993, 0.9995, 0.9994, 0.9992, 0.9991, 0.9993, 0.9992, 0.9994, 0.9995,
        0.9995, 0.9994, 0.9996, 0.9995, 0.9999, 1.0000, 0.9994],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2365,
        0.2304, 0.2260, 0.2226, 0.2202, 0.2183, 0.2169, 0.2159, 0.2151, 0.2145,
        0.2141, 0.2138, 0.2135, 0.2133, 0.2132, 0.2131, 0.2211])
finalReturns:  tensor([1.5017, 1.5098, 1.4889, 1.4507, 1.4019, 1.3464, 1.2867,
1.2243, 1.1600,
        1.0946, 1.0283, 0.9615, 0.8942, 0.8266, 0.7588, 0.6909, 0.6228, 0.5546,
        0.4864, 0.4182, 0.3499, 0.2815, 0.2132, 0.1448, 0.0765])
0,[1e-05,1][1, 10000, 1, 1],1682661769 saved
[378994, 'tensor([0., 0., 1., 0., 0., 0., 0., 0., 0., 0., 0., 0.])',
63858.06679267008, 91921.34476753633, 0.012501846998929977, 1e-05, 1, 0,
'tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9,\n        0])', '[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
```

```
1. 1. 1. 1.\n 1.]', '0,[1e-05,1][1, 10000, 1, 1],1682661769', 25, 50,
153603.58340915042, 166411.3640199337, 63840.19497842904, 134666.6666666667,
131815.9786666667, 116903.27057727946, 116903.27057727946, 134357.5795423984,
134357.5795423984, 74631.0377323958, 116903.27057727946, 134357.5795423984]
policy reset
----------------------------------------
iter  2  stage  24  ep  99999   adversary:  AdversaryModes.constant_95
  actions:  tensor([2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 3, 0, 1, 0, 0, 0, 0,
2, 0, 0, 0,
        0])
loss=  tensor(-0., grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.1469,
0.1469, 0.1469, 0.1469, 0.1469, 0.1469, 0.1469, 0.1469, 0.1469, 0.1469,
        0.1469, 0.1469, 0.1469, 0.1469, 0.1469, 0.1469, 0.1469, 0.1469, 0.1469,
        0.1469, 0.1469, 0.1469, 0.1469, 0.1469, 0.1469, 0.1469]) return=
49505.75711012355
probs of actions:  tensor([0.0101, 0.8958, 0.8695, 0.8334, 0.8086, 0.8203,
0.8380, 0.8371, 0.8358,
        0.8111, 0.8296, 0.8712, 0.0919, 0.0101, 0.8609, 0.0820, 0.8421, 0.8506,
        0.8459, 0.8274, 0.0171, 0.8530, 0.8225, 0.8343, 0.9739],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5108, 0.4048, 0.3274, 0.2747, 0.2382, 0.2126, 0.1943,
0.1811, 0.1716,
        0.1646, 0.1594, 0.1556, 0.1526, 0.1517, 0.1564, 0.1532, 0.1530, 0.1509,
        0.1492, 0.1480, 0.1467, 0.1503, 0.1488, 0.1477, 0.1469])
finalReturns:  tensor([0.])
----------------------------------------
iter  2  stage  23  ep  99999   adversary:  AdversaryModes.constant_95
  actions:  tensor([3, 5, 5, 5, 6, 0, 5, 6, 4, 5, 2, 3, 4, 7, 4, 5, 5, 0, 3, 5,
5, 0, 2, 5,
        0])
loss=  tensor(0.0054, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.3278,
0.3278, 0.3278, 0.3278, 0.3278, 0.3278, 0.3278, 0.3278, 0.3278, 0.3278,
        0.3278, 0.3278, 0.3278, 0.3278, 0.3278, 0.3278, 0.3278, 0.3278, 0.3278,
        0.3278, 0.3278, 0.3278, 0.3278, 0.3278, 0.3278, 0.1610]) return=
55563.48757478318
probs of actions:  tensor([0.0395, 0.4441, 0.3476, 0.4419, 0.0984, 0.2259,
0.3980, 0.0929, 0.1221,
        0.3735, 0.0178, 0.0503, 0.1371, 0.0706, 0.1347, 0.3186, 0.3156, 0.1818,
        0.0519, 0.4018, 0.3445, 0.2042, 0.0182, 0.4977, 0.9952],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5103, 0.4055, 0.3416, 0.2972, 0.2648, 0.2486, 0.2174,
0.2073, 0.2049,
        0.1962, 0.1948, 0.1852, 0.1799, 0.1753, 0.1841, 0.1808, 0.1812, 0.1840,
        0.1728, 0.1698, 0.1729, 0.1777, 0.1687, 0.1643, 0.1712])
finalReturns:  tensor([0.0077, 0.0102])
----------------------------------------
iter  2  stage  22  ep  99999   adversary:  AdversaryModes.constant_95
  actions:  tensor([11,  9, 11, 12, 12,  9, 15, 12, 12,  9, 11, 11, 11, 17, 12,
```

```
11, 12, 11,
         9, 12, 12,  9, 12,  9,  0])
loss=  tensor(0.1197, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.6472,
0.6472, 0.6472, 0.6472, 0.6472, 0.6472, 0.6472, 0.6472, 0.6472,
        0.6472, 0.6472, 0.6472, 0.6472, 0.6472, 0.6472, 0.6472, 0.6472, 0.6472,
        0.6472, 0.6472, 0.6472, 0.6472, 0.6472, 0.4080, 0.1948]) return=
67657.67510325903
probs of actions:  tensor([0.3362, 0.1559, 0.3439, 0.2915, 0.2712, 0.1769,
0.0220, 0.2707, 0.2781,
        0.1934, 0.3567, 0.3823, 0.3320, 0.0054, 0.2852, 0.3265, 0.2812, 0.3373,
        0.1603, 0.2767, 0.2689, 0.1686, 0.4401, 0.2747, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.4259, 0.3619, 0.3234, 0.3002, 0.2896, 0.2550,
0.2639, 0.2567,
        0.2576, 0.2420, 0.2385, 0.2359, 0.2171, 0.2452, 0.2451, 0.2384, 0.2400,
        0.2410, 0.2275, 0.2295, 0.2373, 0.2248, 0.2338, 0.2365])
finalReturns:  tensor([0.0479, 0.0623, 0.0417])
---------------------------------------
iter  2  stage  21  ep  99999   adversary:  AdversaryModes.constant_95
  actions:  tensor([15, 15, 12, 13, 15, 15, 12, 12, 12, 13,  9, 12, 15, 11, 11,
12, 11, 12,
        13, 12, 11, 11, 11,  9,  0])
loss=  tensor(0.6629, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.8528,
0.8528, 0.8528, 0.8528, 0.8528, 0.8528, 0.8528, 0.8528, 0.8528,
        0.8528, 0.8528, 0.8528, 0.8528, 0.8528, 0.8528, 0.8528, 0.8528, 0.8528,
        0.8528, 0.8528, 0.8528, 0.8528, 0.6043, 0.3844, 0.1849]) return=
69577.64726088839
probs of actions:  tensor([0.2457, 0.2400, 0.4711, 0.0916, 0.2515, 0.2377,
0.4269, 0.4518, 0.4405,
        0.0901, 0.0234, 0.4779, 0.2212, 0.1748, 0.1601, 0.4600, 0.1729, 0.4251,
        0.0977, 0.4363, 0.1895, 0.1269, 0.1451, 0.0535, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.4247, 0.3876, 0.3438, 0.3115, 0.2978, 0.2958,
0.2801, 0.2687,
        0.2577, 0.2628, 0.2436, 0.2335, 0.2500, 0.2444, 0.2379, 0.2397, 0.2344,
        0.2322, 0.2374, 0.2393, 0.2364, 0.2343, 0.2367, 0.2387])
finalReturns:  tensor([0.0934, 0.1055, 0.0910, 0.0538])
---------------------------------------
iter  2  stage  20  ep  99999   adversary:  AdversaryModes.constant_95
  actions:  tensor([15, 15, 15, 11, 13, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0408, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.1196,
1.1196, 1.1196, 1.1196, 1.1196, 1.1196, 1.1196, 1.1196, 1.1196,
        1.1196, 1.1196, 1.1196, 1.1196, 1.1196, 1.1196, 1.1196, 1.1196, 1.1196,
        1.1196, 1.1196, 1.1196, 0.8382, 0.5954, 0.3795, 0.1828]) return=
72938.84118085599
probs of actions:  tensor([0.9609, 0.9638, 0.9470, 0.0121, 0.0154, 0.9355,
```

```
0.9524, 0.9420, 0.9463,
        0.9441, 0.9459, 0.9427, 0.9293, 0.9476, 0.9522, 0.9450, 0.9424, 0.9520,
        0.9369, 0.9393, 0.9634, 0.9825, 0.9491, 0.8879, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.4247, 0.3795, 0.3576, 0.3178, 0.2927, 0.2839,
0.2774, 0.2726,
        0.2690, 0.2664, 0.2644, 0.2629, 0.2617, 0.2609, 0.2603, 0.2598, 0.2595,
        0.2592, 0.2590, 0.2588, 0.2587, 0.2587, 0.2586, 0.2810])
finalReturns:  tensor([0.1963, 0.2188, 0.2029, 0.1601, 0.0982])
----------------------------------------
iter  2  stage  19  ep  99999   adversary:  AdversaryModes.constant_95
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0125, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.2935,
1.2935, 1.2935, 1.2935, 1.2935, 1.2935, 1.2935, 1.2935, 1.2935, 1.2935,
        1.2935, 1.2935, 1.2935, 1.2935, 1.2935, 1.2935, 1.2935, 1.2935, 1.2935,
        1.2935, 1.2935, 1.0118, 0.7687, 0.5526, 0.3558, 0.1729]) return=
73445.38232037451
probs of actions:  tensor([0.9926, 0.9932, 0.9889, 0.9853, 0.9862, 0.9855,
0.9897, 0.9877, 0.9884,
        0.9879, 0.9882, 0.9876, 0.9833, 0.9891, 0.9900, 0.9879, 0.9871, 0.9900,
        0.9858, 0.9914, 0.9953, 0.9967, 0.9902, 0.9719, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.4247, 0.3795, 0.3472, 0.3239, 0.3069, 0.2944,
0.2852, 0.2784,
        0.2733, 0.2696, 0.2667, 0.2646, 0.2631, 0.2619, 0.2610, 0.2604, 0.2599,
        0.2595, 0.2592, 0.2590, 0.2589, 0.2587, 0.2587, 0.2811])
finalReturns:  tensor([0.2821, 0.3046, 0.2887, 0.2459, 0.1840, 0.1082])
----------------------------------------
iter  2  stage  18  ep  75370   adversary:  AdversaryModes.constant_95
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 12,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0013, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.4352,
1.4352, 1.4352, 1.4352, 1.4352, 1.4352, 1.4352, 1.4352, 1.4352, 1.4352,
        1.4352, 1.4352, 1.4352, 1.4352, 1.4352, 1.4352, 1.4352, 1.4352, 1.4352,
        1.4352, 1.1611, 0.9233, 0.7110, 0.5169, 0.3359, 0.1645]) return=
73251.7417815017
probs of actions:  tensor([9.9940e-01, 9.9947e-01, 9.9895e-01, 9.9849e-01,
9.9850e-01, 9.9845e-01,
        9.9896e-01, 9.9875e-01, 9.9878e-01, 9.9872e-01, 9.9879e-01, 9.9881e-01,
        9.9813e-01, 9.9887e-01, 9.9903e-01, 9.9874e-01, 9.9859e-01, 6.3129e-04,
        9.9900e-01, 9.9951e-01, 9.9974e-01, 9.9997e-01, 9.9944e-01, 9.9743e-01,
        1.0000e+00], grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.4247, 0.3795, 0.3472, 0.3239, 0.3069, 0.2944,
0.2852, 0.2784,
        0.2733, 0.2696, 0.2667, 0.2646, 0.2631, 0.2619, 0.2610, 0.2604, 0.2680,
```

```
        0.2516, 0.2533, 0.2546, 0.2555, 0.2562, 0.2568, 0.2797])
finalReturns:  tensor([0.3725, 0.3950, 0.3795, 0.3372, 0.2758, 0.2005, 0.1152])
----------------------------------------
iter  2  stage  17  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0020, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.6212,
1.6212, 1.6212, 1.6212, 1.6212, 1.6212, 1.6212, 1.6212, 1.6212,
        1.6212, 1.6212, 1.6212, 1.6212, 1.6212, 1.6212, 1.6212, 1.6212, 1.6212,
        1.3388, 1.0953, 0.8789, 0.6819, 0.4988, 0.3258, 0.1602]) return=
73445.38232037451
probs of actions:  tensor([0.9994, 0.9995, 0.9990, 0.9985, 0.9985, 0.9984,
0.9990, 0.9987, 0.9988,
        0.9987, 0.9988, 0.9988, 0.9981, 0.9989, 0.9990, 0.9987, 0.9986, 0.9990,
        0.9990, 0.9995, 0.9997, 1.0000, 0.9994, 0.9974, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.4247, 0.3795, 0.3472, 0.3239, 0.3069, 0.2944,
0.2852, 0.2784,
        0.2733, 0.2696, 0.2667, 0.2646, 0.2631, 0.2619, 0.2610, 0.2604, 0.2599,
        0.2595, 0.2592, 0.2590, 0.2589, 0.2587, 0.2587, 0.2811])
finalReturns:  tensor([0.4738, 0.4963, 0.4804, 0.4375, 0.3755, 0.2997, 0.2140,
0.1209])
----------------------------------------
iter  2  stage  16  ep  368   adversary:  AdversaryModes.constant_95
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0027, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.7790,
1.7790, 1.7790, 1.7790, 1.7790, 1.7790, 1.7790, 1.7790, 1.7790,
        1.7790, 1.7790, 1.7790, 1.7790, 1.7790, 1.7790, 1.7790, 1.7790, 1.4961,
        1.2522, 1.0356, 0.8384, 0.6552, 0.4821, 0.3164, 0.1562]) return=
73445.38232037451
probs of actions:  tensor([0.9994, 0.9995, 0.9990, 0.9985, 0.9985, 0.9985,
0.9990, 0.9988, 0.9988,
        0.9987, 0.9988, 0.9988, 0.9982, 0.9989, 0.9991, 0.9988, 0.9990, 0.9993,
        0.9990, 0.9995, 0.9998, 1.0000, 0.9995, 0.9975, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.4247, 0.3795, 0.3472, 0.3239, 0.3069, 0.2944,
0.2852, 0.2784,
        0.2733, 0.2696, 0.2667, 0.2646, 0.2631, 0.2619, 0.2610, 0.2604, 0.2599,
        0.2595, 0.2592, 0.2590, 0.2589, 0.2587, 0.2587, 0.2811])
finalReturns:  tensor([0.5764, 0.5989, 0.5829, 0.5401, 0.4780, 0.4022, 0.3164,
0.2234, 0.1249])
----------------------------------------
iter  2  stage  15  ep  147   adversary:  AdversaryModes.constant_95
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
```

```
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0035, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.9344,
1.9344, 1.9344, 1.9344, 1.9344, 1.9344, 1.9344, 1.9344, 1.9344,
        1.9344, 1.9344, 1.9344, 1.9344, 1.9344, 1.9344, 1.9344, 1.6509, 1.4065,
        1.1895, 0.9921, 0.8087, 0.6355, 0.4697, 0.3095, 0.1532]) return=
73445.38232037451
probs of actions:  tensor([0.9994, 0.9995, 0.9990, 0.9985, 0.9986, 0.9985,
0.9990, 0.9988, 0.9988,
        0.9988, 0.9988, 0.9989, 0.9982, 0.9989, 0.9991, 0.9990, 0.9991, 0.9994,
        0.9990, 0.9995, 0.9998, 1.0000, 0.9995, 0.9976, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.4247, 0.3795, 0.3472, 0.3239, 0.3069, 0.2944,
0.2852, 0.2784,
        0.2733, 0.2696, 0.2667, 0.2646, 0.2631, 0.2619, 0.2610, 0.2604, 0.2599,
        0.2595, 0.2592, 0.2590, 0.2589, 0.2587, 0.2587, 0.2811])
finalReturns:  tensor([0.6820, 0.7045, 0.6885, 0.6456, 0.5835, 0.5077, 0.4219,
0.3288, 0.2303,
        0.1279])
----------------------------------------
iter  2  stage  14  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0046, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.0884,
2.0884, 2.0884, 2.0884, 2.0884, 2.0884, 2.0884, 2.0884, 2.0884,
        2.0884, 2.0884, 2.0884, 2.0884, 2.0884, 2.0884, 1.8040, 1.5590, 1.3416,
        1.1439, 0.9602, 0.7869, 0.6210, 0.4606, 0.3043, 0.1510]) return=
73445.38232037451
probs of actions:  tensor([0.9994, 0.9995, 0.9990, 0.9985, 0.9986, 0.9985,
0.9990, 0.9988, 0.9988,
        0.9988, 0.9988, 0.9989, 0.9982, 0.9989, 0.9991, 0.9990, 0.9991, 0.9994,
        0.9990, 0.9995, 0.9998, 1.0000, 0.9995, 0.9976, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.4247, 0.3795, 0.3472, 0.3239, 0.3069, 0.2944,
0.2852, 0.2784,
        0.2733, 0.2696, 0.2667, 0.2646, 0.2631, 0.2619, 0.2610, 0.2604, 0.2599,
        0.2595, 0.2592, 0.2590, 0.2589, 0.2587, 0.2587, 0.2811])
finalReturns:  tensor([0.7900, 0.8125, 0.7964, 0.7534, 0.6913, 0.6154, 0.5295,
0.4364, 0.3379,
        0.2354, 0.1301])
----------------------------------------
iter  2  stage  13  ep  58   adversary:  AdversaryModes.constant_95
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0058, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.2417,
2.2417, 2.2417, 2.2417, 2.2417, 2.2417, 2.2417, 2.2417, 2.2417,
        2.2417, 2.2417, 2.2417, 2.2417, 2.2417, 2.2417, 1.9561, 1.7103, 1.4924, 1.2942,
```

```
        1.1103, 0.9367, 0.7707, 0.6102, 0.4538, 0.3005, 0.1494]) return=
73445.38232037451
probs of actions:  tensor([0.9994, 0.9995, 0.9990, 0.9985, 0.9985, 0.9985,
0.9990, 0.9988, 0.9988,
        0.9987, 0.9988, 0.9988, 0.9982, 0.9990, 0.9991, 0.9990, 0.9991, 0.9994,
        0.9990, 0.9995, 0.9998, 1.0000, 0.9994, 0.9975, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.4247, 0.3795, 0.3472, 0.3239, 0.3069, 0.2944,
0.2852, 0.2784,
        0.2733, 0.2696, 0.2667, 0.2646, 0.2631, 0.2619, 0.2610, 0.2604, 0.2599,
        0.2595, 0.2592, 0.2590, 0.2589, 0.2587, 0.2587, 0.2811])
finalReturns:  tensor([0.8997, 0.9222, 0.9061, 0.8630, 0.8008, 0.7248, 0.6389,
0.5457, 0.4472,
        0.3447, 0.2393, 0.1317])
----------------------------------------
iter  2  stage  12  ep  8390   adversary:  AdversaryModes.constant_95
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0049, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.3953,
2.3953, 2.3953, 2.3953, 2.3953, 2.3953, 2.3953, 2.3953, 2.3953,
        2.3953, 2.3953, 2.3953, 2.3953, 2.1081, 1.8612, 1.6425, 1.4438, 1.2595,
        1.0856, 0.9193, 0.7587, 0.6022, 0.4488, 0.2976, 0.1482]) return=
73445.38232037451
probs of actions:  tensor([0.9996, 0.9996, 0.9992, 0.9988, 0.9989, 0.9988,
0.9992, 0.9990, 0.9990,
        0.9990, 0.9991, 0.9991, 0.9990, 0.9993, 0.9995, 0.9994, 0.9994, 0.9999,
        0.9993, 0.9997, 0.9999, 1.0000, 0.9996, 0.9979, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.4247, 0.3795, 0.3472, 0.3239, 0.3069, 0.2944,
0.2852, 0.2784,
        0.2733, 0.2696, 0.2667, 0.2646, 0.2631, 0.2619, 0.2610, 0.2604, 0.2599,
        0.2595, 0.2592, 0.2590, 0.2589, 0.2587, 0.2587, 0.2811])
finalReturns:  tensor([1.0108, 1.0333, 1.0171, 0.9739, 0.9116, 0.8355, 0.7495,
0.6563, 0.5577,
        0.4552, 0.3497, 0.2421, 0.1329])
----------------------------------------
iter  2  stage  11  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0062, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.5497,
2.5497, 2.5497, 2.5497, 2.5497, 2.5497, 2.5497, 2.5497, 2.5497,
        2.5497, 2.5497, 2.5497, 2.2605, 2.0121, 1.7924, 1.5929, 1.4081, 1.2338,
        1.0673, 0.9064, 0.7498, 0.5962, 0.4450, 0.2955, 0.1473]) return=
73445.38232037451
probs of actions:  tensor([0.9996, 0.9996, 0.9992, 0.9988, 0.9989, 0.9988,
0.9992, 0.9990, 0.9990,
```

```
        0.9990, 0.9991, 0.9991, 0.9990, 0.9993, 0.9995, 0.9994, 0.9994, 0.9999,
        0.9993, 0.9997, 0.9999, 1.0000, 0.9996, 0.9979, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.4247, 0.3795, 0.3472, 0.3239, 0.3069, 0.2944,
0.2852, 0.2784,
        0.2733, 0.2696, 0.2667, 0.2646, 0.2631, 0.2619, 0.2610, 0.2604, 0.2599,
        0.2595, 0.2592, 0.2590, 0.2589, 0.2587, 0.2587, 0.2811])
finalReturns:  tensor([1.1231, 1.1456, 1.1293, 1.0859, 1.0235, 0.9473, 0.8612,
0.7678, 0.6692,
        0.5666, 0.4611, 0.3535, 0.2443, 0.1338])
---------------------------------------
iter  2  stage  10  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0078, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.7060,
2.7060, 2.7060, 2.7060, 2.7060, 2.7060, 2.7060, 2.7060, 2.7060,
        2.7060, 2.7060, 2.4139, 2.1636, 1.9425, 1.7421, 1.5565, 1.3817, 1.2148,
        1.0537, 0.8968, 0.7431, 0.5918, 0.4422, 0.2939, 0.1466]) return=
73445.38232037451
probs of actions:  tensor([0.9996, 0.9996, 0.9992, 0.9988, 0.9989, 0.9988,
0.9992, 0.9990, 0.9990,
        0.9990, 0.9991, 0.9991, 0.9990, 0.9993, 0.9995, 0.9994, 0.9994, 0.9999,
        0.9993, 0.9997, 0.9999, 1.0000, 0.9996, 0.9979, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.4247, 0.3795, 0.3472, 0.3239, 0.3069, 0.2944,
0.2852, 0.2784,
        0.2733, 0.2696, 0.2667, 0.2646, 0.2631, 0.2619, 0.2610, 0.2604, 0.2599,
        0.2595, 0.2592, 0.2590, 0.2589, 0.2587, 0.2587, 0.2811])
finalReturns:  tensor([1.2364, 1.2589, 1.2424, 1.1989, 1.1362, 1.0599, 0.9736,
0.8802, 0.7815,
        0.6788, 0.5733, 0.4656, 0.3563, 0.2459, 0.1345])
---------------------------------------
iter  2  stage  9  ep  3   adversary:  AdversaryModes.constant_95
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0095, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.8650,
2.8650, 2.8650, 2.8650, 2.8650, 2.8650, 2.8650, 2.8650, 2.8650,
        2.8650, 2.5692, 2.3163, 2.0933, 1.8916, 1.7051, 1.5296, 1.3622, 1.2007,
        1.0436, 0.8897, 0.7382, 0.5884, 0.4401, 0.2927, 0.1461]) return=
73445.38232037451
probs of actions:  tensor([0.9996, 0.9996, 0.9992, 0.9988, 0.9989, 0.9988,
0.9992, 0.9990, 0.9991,
        0.9990, 0.9991, 0.9991, 0.9990, 0.9993, 0.9995, 0.9994, 0.9994, 0.9999,
        0.9993, 0.9997, 0.9999, 1.0000, 0.9996, 0.9979, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.4247, 0.3795, 0.3472, 0.3239, 0.3069, 0.2944,
```

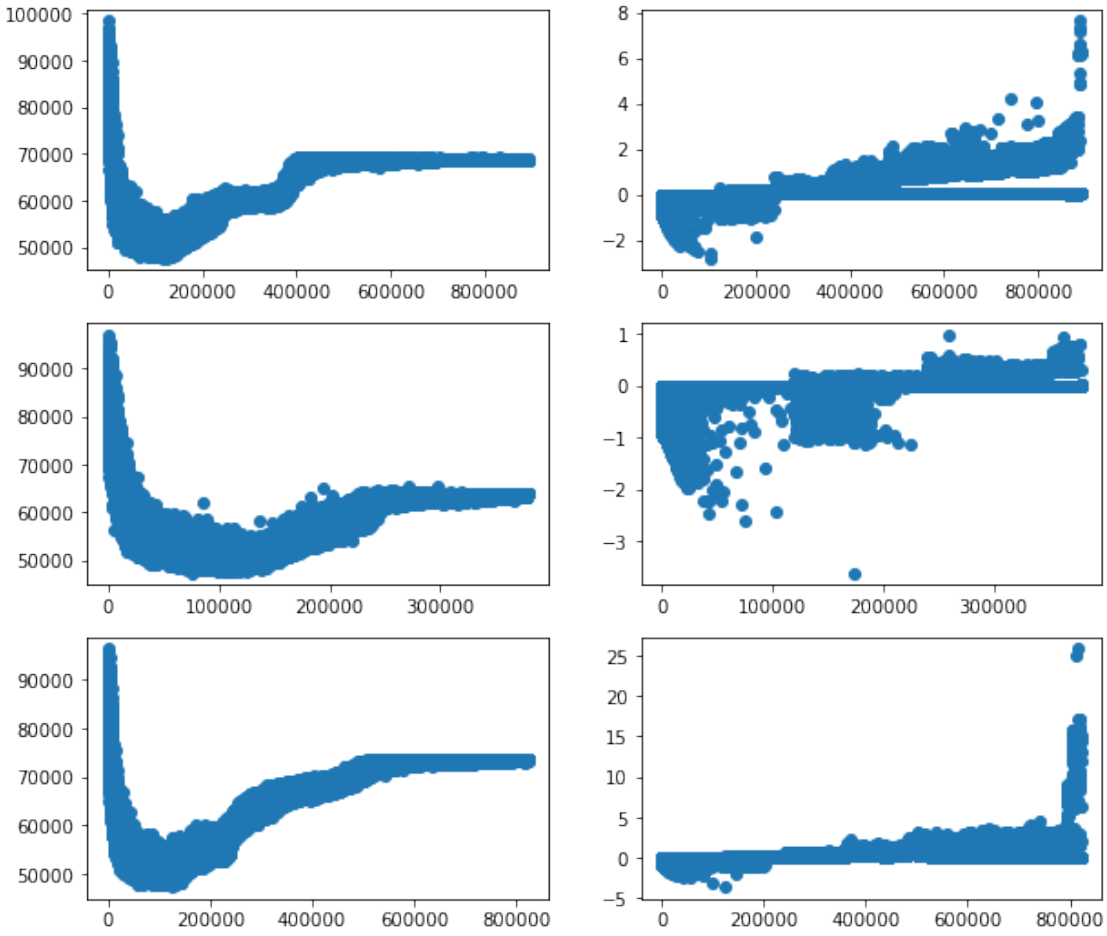
```
0.2852, 0.2784,
        0.2733, 0.2696, 0.2667, 0.2646, 0.2631, 0.2619, 0.2610, 0.2604, 0.2599,
        0.2595, 0.2592, 0.2590, 0.2589, 0.2587, 0.2587, 0.2811])
finalReturns:  tensor([1.3506, 1.3731, 1.3565, 1.3127, 1.2498, 1.1732, 1.0868,
0.9932, 0.8943,
        0.7916, 0.6860, 0.5782, 0.4689, 0.3584, 0.2471, 0.1350])
----------------------------------------
iter  2  stage  8  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0113, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.0282,
3.0282, 3.0282, 3.0282, 3.0282, 3.0282, 3.0282, 3.0282, 3.0282,
        2.7273, 2.4709, 2.2454, 2.0419, 1.8542, 1.6778, 1.5097, 1.3477, 1.1902,
        1.0360, 0.8843, 0.7344, 0.5860, 0.4385, 0.2918, 0.1457]) return=
73445.38232037451
probs of actions:  tensor([0.9996, 0.9996, 0.9992, 0.9988, 0.9989, 0.9988,
0.9992, 0.9990, 0.9991,
        0.9990, 0.9991, 0.9991, 0.9990, 0.9993, 0.9995, 0.9994, 0.9994, 0.9999,
        0.9993, 0.9997, 0.9999, 1.0000, 0.9996, 0.9979, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.4247, 0.3795, 0.3472, 0.3239, 0.3069, 0.2944,
0.2852, 0.2784,
        0.2733, 0.2696, 0.2667, 0.2646, 0.2631, 0.2619, 0.2610, 0.2604, 0.2599,
        0.2595, 0.2592, 0.2590, 0.2589, 0.2587, 0.2587, 0.2811])
finalReturns:  tensor([1.4659, 1.4884, 1.4715, 1.4274, 1.3641, 1.2872, 1.2005,
1.1067, 1.0077,
        0.9048, 0.7991, 0.6913, 0.5820, 0.4714, 0.3600, 0.2480, 0.1354])
----------------------------------------
iter  2  stage  7  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0132, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.1970,
3.1970, 3.1970, 3.1970, 3.1970, 3.1970, 3.1970, 3.1970, 2.8893,
        2.6282, 2.3995, 2.1936, 2.0042, 1.8265, 1.6575, 1.4949, 1.3369, 1.1823,
        1.0304, 0.8803, 0.7317, 0.5841, 0.4373, 0.2911, 0.1454]) return=
73445.38232037451
probs of actions:  tensor([0.9996, 0.9996, 0.9992, 0.9988, 0.9989, 0.9988,
0.9992, 0.9990, 0.9991,
        0.9990, 0.9991, 0.9991, 0.9990, 0.9993, 0.9995, 0.9994, 0.9994, 0.9999,
        0.9993, 0.9997, 0.9999, 1.0000, 0.9996, 0.9979, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.4247, 0.3795, 0.3472, 0.3239, 0.3069, 0.2944,
0.2852, 0.2784,
        0.2733, 0.2696, 0.2667, 0.2646, 0.2631, 0.2619, 0.2610, 0.2604, 0.2599,
        0.2595, 0.2592, 0.2590, 0.2589, 0.2587, 0.2587, 0.2811])
finalReturns:  tensor([1.5822, 1.6047, 1.5874, 1.5429, 1.4792, 1.4019, 1.3149,
```

```
1.2208, 1.1215,
        1.0185, 0.9127, 0.8048, 0.6953, 0.5847, 0.4733, 0.3612, 0.2486, 0.1357])
----------------------------------------
iter  2  stage  6  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0150, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.3738,
3.3738, 3.3738, 3.3738, 3.3738, 3.3738, 3.3738, 3.0569, 2.7895,
        2.5563, 2.3473, 2.1555, 1.9762, 1.8060, 1.6425, 1.4838, 1.3288, 1.1765,
        1.0261, 0.8773, 0.7296, 0.5827, 0.4364, 0.2906, 0.1452]) return=
73445.38232037451
probs of actions:  tensor([0.9996, 0.9996, 0.9992, 0.9988, 0.9989, 0.9988,
0.9992, 0.9990, 0.9991,
        0.9990, 0.9991, 0.9991, 0.9990, 0.9993, 0.9995, 0.9994, 0.9994, 0.9999,
        0.9993, 0.9997, 0.9999, 1.0000, 0.9996, 0.9979, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.4247, 0.3795, 0.3472, 0.3239, 0.3069, 0.2944,
0.2852, 0.2784,
        0.2733, 0.2696, 0.2667, 0.2646, 0.2631, 0.2619, 0.2610, 0.2604, 0.2599,
        0.2595, 0.2592, 0.2590, 0.2589, 0.2587, 0.2587, 0.2811])
finalReturns:  tensor([1.6998, 1.7223, 1.7046, 1.6594, 1.5951, 1.5173, 1.4298,
1.3353, 1.2358,
        1.1326, 1.0266, 0.9185, 0.8090, 0.6983, 0.5868, 0.4747, 0.3621, 0.2491,
        0.1359])
----------------------------------------
iter  2  stage  5  ep  18525   adversary:  AdversaryModes.constant_95
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0105, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.5615,
3.5615, 3.5615, 3.5615, 3.5615, 3.5615, 3.2322, 2.9561, 2.7169,
        2.5036, 2.3089, 2.1274, 1.9555, 1.7908, 1.6313, 1.4756, 1.3228, 1.1721,
        1.0230, 0.8751, 0.7280, 0.5816, 0.4358, 0.2903, 0.1450]) return=
73445.38232037451
probs of actions:  tensor([0.9996, 0.9996, 0.9993, 0.9990, 0.9990, 0.9990,
0.9993, 0.9997, 0.9994,
        0.9992, 1.0000, 0.9999, 0.9993, 0.9994, 0.9995, 0.9996, 0.9997, 1.0000,
        0.9995, 1.0000, 1.0000, 1.0000, 0.9997, 0.9984, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.4247, 0.3795, 0.3472, 0.3239, 0.3069, 0.2944,
0.2852, 0.2784,
        0.2733, 0.2696, 0.2667, 0.2646, 0.2631, 0.2619, 0.2610, 0.2604, 0.2599,
        0.2595, 0.2592, 0.2590, 0.2589, 0.2587, 0.2587, 0.2811])
finalReturns:  tensor([1.8189, 1.8414, 1.8231, 1.7771, 1.7120, 1.6335, 1.5454,
1.4505, 1.3506,
        1.2470, 1.1408, 1.0326, 0.9229, 0.8122, 0.7006, 0.5884, 0.4757, 0.3627,
        0.2495, 0.1361])
```

```
---------------------------------------
iter  2  stage  4  ep  15   adversary:  AdversaryModes.constant_95
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0127, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.7642,
3.7642, 3.7642, 3.7642, 3.7642, 3.4179, 3.1301, 2.8828, 2.6638,
        2.4649, 2.2805, 2.1065, 1.9402, 1.7795, 1.6229, 1.4694, 1.3183, 1.1688,
        1.0206, 0.8734, 0.7269, 0.5809, 0.4353, 0.2900, 0.1449]) return=
73445.38232037451
probs of actions:  tensor([0.9996, 0.9996, 0.9993, 0.9990, 0.9990, 0.9990,
0.9993, 0.9997, 0.9994,
        0.9992, 1.0000, 0.9999, 0.9993, 0.9994, 0.9995, 0.9996, 0.9997, 1.0000,
        0.9995, 1.0000, 1.0000, 1.0000, 0.9997, 0.9984, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.4247, 0.3795, 0.3472, 0.3239, 0.3069, 0.2944,
0.2852, 0.2784,
        0.2733, 0.2696, 0.2667, 0.2646, 0.2631, 0.2619, 0.2610, 0.2604, 0.2599,
        0.2595, 0.2592, 0.2590, 0.2589, 0.2587, 0.2587, 0.2811])
finalReturns:  tensor([1.9401, 1.9626, 1.9435, 1.8964, 1.8302, 1.7507, 1.6619,
1.5663, 1.4659,
        1.3619, 1.2554, 1.1470, 1.0371, 0.9262, 0.8145, 0.7022, 0.5895, 0.4765,
        0.3632, 0.2498, 0.1362])
---------------------------------------
iter  2  stage  3  ep  7   adversary:  AdversaryModes.constant_95
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0151, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.9877,
3.9877, 3.9877, 3.9877, 3.6180, 3.3143, 3.0560, 2.8292, 2.6248,
        2.4363, 2.2594, 2.0910, 1.9287, 1.7710, 1.6167, 1.4648, 1.3149, 1.1664,
        1.0188, 0.8721, 0.7260, 0.5803, 0.4349, 0.2898, 0.1448]) return=
73445.38232037451
probs of actions:  tensor([0.9996, 0.9996, 0.9993, 0.9990, 0.9990, 0.9990,
0.9993, 0.9997, 0.9994,
        0.9992, 1.0000, 0.9999, 0.9993, 0.9994, 0.9995, 0.9996, 0.9997, 1.0000,
        0.9995, 1.0000, 1.0000, 1.0000, 0.9997, 0.9984, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.4247, 0.3795, 0.3472, 0.3239, 0.3069, 0.2944,
0.2852, 0.2784,
        0.2733, 0.2696, 0.2667, 0.2646, 0.2631, 0.2619, 0.2610, 0.2604, 0.2599,
        0.2595, 0.2592, 0.2590, 0.2589, 0.2587, 0.2587, 0.2811])
finalReturns:  tensor([2.0639, 2.0864, 2.0661, 2.0176, 1.9500, 1.8693, 1.7793,
1.6829, 1.5818,
        1.4773, 1.3704, 1.2617, 1.1516, 1.0405, 0.9287, 0.8163, 0.7035, 0.5904,
        0.4771, 0.3636, 0.2500, 0.1363])
---------------------------------------
iter  2  stage  2  ep  0   adversary:  AdversaryModes.constant_95
```

```
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0170, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.2401,
4.2401, 4.2401, 3.8380, 3.5125, 3.2391, 3.0017, 2.7898, 2.5959,
        2.4151, 2.2438, 2.0794, 1.9201, 1.7646, 1.6120, 1.4614, 1.3124, 1.1645,
        1.0175, 0.8712, 0.7253, 0.5798, 0.4346, 0.2896, 0.1447]) return=
73445.38232037451
probs of actions:  tensor([0.9996, 0.9996, 0.9993, 0.9990, 0.9990, 0.9990,
0.9993, 0.9997, 0.9994,
        0.9992, 1.0000, 0.9999, 0.9993, 0.9994, 0.9995, 0.9996, 0.9997, 1.0000,
        0.9995, 1.0000, 1.0000, 1.0000, 0.9997, 0.9984, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.4247, 0.3795, 0.3472, 0.3239, 0.3069, 0.2944,
0.2852, 0.2784,
        0.2733, 0.2696, 0.2667, 0.2646, 0.2631, 0.2619, 0.2610, 0.2604, 0.2599,
        0.2595, 0.2592, 0.2590, 0.2589, 0.2587, 0.2587, 0.2811])
finalReturns:  tensor([2.1910, 2.2135, 2.1918, 2.1414, 2.0719, 1.9895, 1.8981,
1.8006, 1.6986,
        1.5934, 1.4859, 1.3767, 1.2663, 1.1550, 1.0430, 0.9305, 0.8176, 0.7044,
        0.5911, 0.4775, 0.3639, 0.2502, 0.1364])
----------------------------------------
iter  2  stage  1  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0182, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.5331,
4.5331, 4.0859, 3.7300, 3.4358, 3.1839, 2.9618, 2.7606, 2.5745,
        2.3993, 2.2321, 2.0708, 1.9137, 1.7599, 1.6085, 1.4588, 1.3105, 1.1631,
        1.0165, 0.8705, 0.7248, 0.5795, 0.4344, 0.2895, 0.1447]) return=
73445.38232037451
probs of actions:  tensor([0.9996, 0.9996, 0.9993, 0.9990, 0.9990, 0.9990,
0.9993, 0.9997, 0.9994,
        0.9992, 1.0000, 0.9999, 0.9993, 0.9994, 0.9995, 0.9996, 0.9997, 1.0000,
        0.9995, 1.0000, 1.0000, 1.0000, 0.9997, 0.9984, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.4247, 0.3795, 0.3472, 0.3239, 0.3069, 0.2944,
0.2852, 0.2784,
        0.2733, 0.2696, 0.2667, 0.2646, 0.2631, 0.2619, 0.2610, 0.2604, 0.2599,
        0.2595, 0.2592, 0.2590, 0.2589, 0.2587, 0.2587, 0.2811])
finalReturns:  tensor([2.3227, 2.3452, 2.3216, 2.2686, 2.1965, 2.1118, 2.0186,
1.9195, 1.8163,
        1.7102, 1.6020, 1.4923, 1.3815, 1.2698, 1.1576, 1.0449, 0.9319, 0.8186,
        0.7052, 0.5916, 0.4779, 0.3641, 0.2503, 0.1364])
----------------------------------------
iter  2  stage  0  ep  0   adversary:  AdversaryModes.constant_95
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
```

```
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0196, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.8841,
4.3729, 3.9744, 3.6513, 3.3795, 3.1433, 2.9322, 2.7389, 2.5586,
        2.3875, 2.2234, 2.0643, 1.9089, 1.7564, 1.6059, 1.4569, 1.3091, 1.1621,
        1.0158, 0.8699, 0.7245, 0.5793, 0.4343, 0.2894, 0.1447]) return=
73445.38232037451
probs of actions:  tensor([0.9996, 0.9996, 0.9993, 0.9990, 0.9990, 0.9990,
0.9993, 0.9997, 0.9994,
        0.9992, 1.0000, 0.9999, 0.9993, 0.9994, 0.9996, 0.9996, 0.9997, 1.0000,
        0.9995, 1.0000, 1.0000, 1.0000, 0.9997, 0.9984, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.4247, 0.3795, 0.3472, 0.3239, 0.3069, 0.2944,
0.2852, 0.2784,
        0.2733, 0.2696, 0.2667, 0.2646, 0.2631, 0.2619, 0.2610, 0.2604, 0.2599,
        0.2595, 0.2592, 0.2590, 0.2589, 0.2587, 0.2587, 0.2811])
finalReturns:  tensor([2.4604, 2.4829, 2.4566, 2.4002, 2.3248, 2.2371, 2.1414,
2.0403, 1.9355,
        1.8281, 1.7189, 1.6085, 1.4971, 1.3850, 1.2725, 1.1595, 1.0463, 0.9329,
        0.8194, 0.7057, 0.5919, 0.4781, 0.3643, 0.2504, 0.1364])
0,[1e-05,1][1, 10000, 1, 1],1682682010 saved
[822902, 'tensor([0., 0., 1., 0., 0., 0., 0., 0., 0., 0., 0., 0.])',
73445.38232037451, 85129.48297577976, 0.01956232078373432, 1e-05, 1, 0,
'tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15,\n        15, 15, 15, 15, 15, 15,  0])', '[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.\n 1.]', '0,[1e-05,1][1, 10000, 1,
1],1682682010', 25, 50, 161287.58344151574, 184572.3091501231,
73445.38232037451, 135544.55463053397, 132647.75466666667, 112524.66365021843,
112524.66365021843, 117051.59338141435, 117051.59338141435, 85375.85726043607,
112524.66365021843, 117051.59338141435]
```




```
policy reset
----------------------------------------
iter  0  stage  24  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([ 1,  1,  0, 28,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,
0,  0,  0,
         0,  0,  0,  0,  1,  0,  0])
loss=  tensor(-0., grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.5204,
0.5204, 0.5204, 0.5204, 0.5204, 0.5204, 0.5204, 0.5204, 0.5204,
        0.5204, 0.5204, 0.5204, 0.5204, 0.5204, 0.5204, 0.5204, 0.5204, 0.5204,
        0.5204, 0.5204, 0.5204, 0.5204, 0.5204, 0.5204, 0.5204]) return=
130427.71758286069
probs of actions:  tensor([1.2898e-01, 1.4863e-01, 7.8200e-01, 5.0099e-04,
8.0009e-01, 8.2463e-01,
        7.8552e-01, 7.7865e-01, 8.1916e-01, 7.8481e-01, 7.8624e-01, 7.9122e-01,
        8.0141e-01, 1.2257e-01, 8.0664e-01, 8.0212e-01, 8.3226e-01, 8.0869e-01,
        8.1746e-01, 8.0829e-01, 8.0075e-01, 8.0096e-01, 1.2746e-01, 8.1773e-01,
        9.4606e-01], grad_fn=<ExpBackward0>)
rewards:  tensor([0.5111, 0.5273, 0.5234, 0.4424, 0.6274, 0.4963, 0.5276,
```

```
0.5197, 0.5217,
        0.5212, 0.5213, 0.5213, 0.5213, 0.5212, 0.5249, 0.5204, 0.5215, 0.5212,
        0.5213, 0.5213, 0.5213, 0.5213, 0.5212, 0.5249, 0.5204])
finalReturns:  tensor([0.])
----------------------------------------
iter  0  stage  23  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([16,  0,  1, 16,  0, 16,  0, 16,  0,  0,  4, 18,  0,  0, 16,
7, 13, 18,
        16, 16, 16,  0, 16, 16,  0])
loss=  tensor(0.0165, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.0899,
1.0899, 1.0899, 1.0899, 1.0899, 1.0899, 1.0899, 1.0899, 1.0899,
        1.0899, 1.0899, 1.0899, 1.0899, 1.0899, 1.0899, 1.0899, 1.0899, 1.0899,
        1.0899, 1.0899, 1.0899, 1.0899, 1.0899, 1.0899, 0.5062]) return=
133455.20582218055
probs of actions:  tensor([0.3193, 0.3793, 0.0503, 0.2874, 0.5107, 0.3032,
0.3821, 0.2562, 0.4683,
        0.3529, 0.0051, 0.1597, 0.2929, 0.4627, 0.3191, 0.0089, 0.0342, 0.1184,
        0.2100, 0.3366, 0.3129, 0.4630, 0.3278, 0.6132, 0.9927],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4856, 0.5833, 0.5062, 0.5031, 0.5787, 0.4818, 0.5843,
0.4805, 0.5847,
        0.5060, 0.5235, 0.5024, 0.5847, 0.5060, 0.4995, 0.5747, 0.5155, 0.5340,
        0.5510, 0.5409, 0.5434, 0.5684, 0.4842, 0.5581, 0.5648])
finalReturns:  tensor([0.0329, 0.0585])
----------------------------------------
iter  0  stage  22  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([18, 18, 18, 18, 18, 16, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 16,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0506, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.5443,
1.5443, 1.5443, 1.5443, 1.5443, 1.5443, 1.5443, 1.5443, 1.5443,
        1.5443, 1.5443, 1.5443, 1.5443, 1.5443, 1.5443, 1.5443, 1.5443, 1.5443,
        1.5443, 1.5443, 1.5443, 1.5443, 1.5443, 0.9698, 0.4614]) return=
135378.46941865046
probs of actions:  tensor([0.7424, 0.7524, 0.7040, 0.7659, 0.7584, 0.2114,
0.7666, 0.7316, 0.7425,
        0.7602, 0.7405, 0.7840, 0.7915, 0.7560, 0.7886, 0.7388, 0.7687, 0.1881,
        0.7565, 0.7638, 0.7718, 0.7267, 0.8800, 0.7854, 0.9964],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5381, 0.5432, 0.5419, 0.5490, 0.5346,
0.5441, 0.5417,
        0.5423, 0.5421, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5490,
        0.5346, 0.5441, 0.5417, 0.5423, 0.5421, 0.5422, 0.5746])
finalReturns:  tensor([0.1146, 0.1470, 0.1132])
----------------------------------------
iter  0  stage  21  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
```

```
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0087, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.9719,
1.9719, 1.9719, 1.9719, 1.9719, 1.9719, 1.9719, 1.9719, 1.9719,
        1.9719, 1.9719, 1.9719, 1.9719, 1.9719, 1.9719, 1.9719, 1.9719, 1.9719,
        1.9719, 1.9719, 1.9719, 1.9719, 1.3974, 0.8890, 0.4276]) return=
135363.21866666665
probs of actions:  tensor([0.9761, 0.9773, 0.9695, 0.9786, 0.9814, 0.9793,
0.9797, 0.9766, 0.9777,
        0.9793, 0.9776, 0.9815, 0.9828, 0.9794, 0.9828, 0.9772, 0.9826, 0.9781,
        0.9808, 0.9789, 0.9796, 0.9874, 0.9920, 0.9846, 0.9985],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5381, 0.5432, 0.5419, 0.5422, 0.5421,
0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5746])
finalReturns:  tensor([0.2291, 0.2615, 0.2277, 0.1469])
----------------------------------------
iter  0  stage  20  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0133, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.3751,
2.3751, 2.3751, 2.3751, 2.3751, 2.3751, 2.3751, 2.3751, 2.3751,
        2.3751, 2.3751, 2.3751, 2.3751, 2.3751, 2.3751, 2.3751, 2.3751, 2.3751,
        2.3751, 2.3751, 2.3751, 1.8005, 1.2922, 0.8308, 0.4032]) return=
135363.21866666665
probs of actions:  tensor([0.9824, 0.9830, 0.9781, 0.9842, 0.9869, 0.9853,
0.9850, 0.9833, 0.9834,
        0.9845, 0.9833, 0.9865, 0.9871, 0.9851, 0.9869, 0.9831, 0.9873, 0.9838,
        0.9868, 0.9844, 0.9890, 0.9896, 0.9933, 0.9911, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5381, 0.5432, 0.5419, 0.5422, 0.5421,
0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5746])
finalReturns:  tensor([0.3681, 0.4005, 0.3667, 0.2859, 0.1714])
----------------------------------------
iter  0  stage  19  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0049, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.7604,
2.7604, 2.7604, 2.7604, 2.7604, 2.7604, 2.7604, 2.7604, 2.7604,
        2.7604, 2.7604, 2.7604, 2.7604, 2.7604, 2.7604, 2.7604, 2.7604, 2.7604,
        2.7604, 2.7604, 2.1858, 1.6775, 1.2161, 0.7885, 0.3853]) return=
135363.21866666665
probs of actions:  tensor([0.9953, 0.9953, 0.9940, 0.9958, 0.9968, 0.9965,
0.9960, 0.9958, 0.9957,
```

```
        0.9960, 0.9956, 0.9964, 0.9968, 0.9963, 0.9966, 0.9957, 0.9969, 0.9957,
        0.9966, 0.9966, 0.9979, 0.9984, 0.9988, 0.9983, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5381, 0.5432, 0.5419, 0.5422, 0.5421,
0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5746])
finalReturns:  tensor([0.5250, 0.5574, 0.5236, 0.4428, 0.3283, 0.1893])
----------------------------------------
iter  0  stage  18  ep  27994   adversary:  AdversaryModes.imitation_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0030, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.1325,
3.1325, 3.1325, 3.1325, 3.1325, 3.1325, 3.1325, 3.1325, 3.1325,
        3.1325, 3.1325, 3.1325, 3.1325, 3.1325, 3.1325, 3.1325, 3.1325, 3.1325,
        3.1325, 2.5580, 2.0496, 1.5882, 1.1606, 0.7574, 0.3721]) return=
135363.21866666665
probs of actions:  tensor([0.9977, 0.9977, 0.9971, 0.9980, 0.9985, 0.9984,
0.9980, 0.9980, 0.9980,
        0.9981, 0.9979, 0.9982, 0.9985, 0.9983, 0.9984, 0.9980, 0.9986, 0.9980,
        0.9990, 0.9989, 0.9989, 0.9993, 0.9996, 0.9995, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5381, 0.5432, 0.5419, 0.5422, 0.5421,
0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5746])
finalReturns:  tensor([0.6950, 0.7274, 0.6936, 0.6128, 0.4983, 0.3593, 0.2024])
----------------------------------------
iter  0  stage  17  ep  19492   adversary:  AdversaryModes.imitation_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0024, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.4950,
3.4950, 3.4950, 3.4950, 3.4950, 3.4950, 3.4950, 3.4950, 3.4950,
        3.4950, 3.4950, 3.4950, 3.4950, 3.4950, 3.4950, 3.4950, 3.4950, 3.4950,
        2.9204, 2.4120, 1.9507, 1.5230, 1.1199, 0.7346, 0.3624]) return=
135363.21866666665
probs of actions:  tensor([0.9987, 0.9987, 0.9983, 0.9988, 0.9992, 0.9991,
0.9989, 0.9988, 0.9989,
        0.9989, 0.9988, 0.9990, 0.9992, 0.9990, 0.9991, 0.9988, 0.9992, 0.9990,
        0.9997, 0.9995, 0.9996, 0.9996, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5381, 0.5432, 0.5419, 0.5422, 0.5421,
0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5746])
finalReturns:  tensor([0.8747, 0.9071, 0.8733, 0.7926, 0.6780, 0.5390, 0.3821,
```

```
0.2121])
---------------------------------------
iter  0  stage  16  ep  0   adversary:  AdversaryModes.imitation_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0036, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.8502,
3.8502, 3.8502, 3.8502, 3.8502, 3.8502, 3.8502, 3.8502, 3.8502, 3.8502,
        3.8502, 3.8502, 3.8502, 3.8502, 3.8502, 3.8502, 3.8502, 3.8502, 3.2757,
        2.7673, 2.3059, 1.8783, 1.4751, 1.0898, 0.7177, 0.3552]) return=
135363.21866666665
probs of actions:  tensor([0.9987, 0.9987, 0.9983, 0.9988, 0.9992, 0.9991,
0.9989, 0.9988, 0.9989,
        0.9989, 0.9988, 0.9990, 0.9992, 0.9990, 0.9991, 0.9989, 0.9992, 0.9990,
        0.9997, 0.9995, 0.9996, 0.9996, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5381, 0.5432, 0.5419, 0.5422, 0.5421,
0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5746])
finalReturns:  tensor([1.0617, 1.0941, 1.0603, 0.9795, 0.8649, 0.7259, 0.5691,
0.3991, 0.2193])
---------------------------------------
iter  0  stage  15  ep  79   adversary:  AdversaryModes.imitation_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0051, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.2001,
4.2001, 4.2001, 4.2001, 4.2001, 4.2001, 4.2001, 4.2001, 4.2001, 4.2001,
        4.2001, 4.2001, 4.2001, 4.2001, 4.2001, 4.2001, 4.2001, 3.6255, 3.1172,
        2.6558, 2.2282, 1.8250, 1.4397, 1.0676, 0.7051, 0.3499]) return=
135363.21866666665
probs of actions:  tensor([0.9988, 0.9987, 0.9984, 0.9989, 0.9992, 0.9991,
0.9989, 0.9989, 0.9989,
        0.9990, 0.9988, 0.9990, 0.9992, 0.9991, 0.9991, 0.9990, 0.9993, 0.9990,
        0.9997, 0.9995, 0.9996, 0.9997, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5381, 0.5432, 0.5419, 0.5422, 0.5421,
0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5746])
finalReturns:  tensor([1.2539, 1.2863, 1.2525, 1.1718, 1.0572, 0.9182, 0.7613,
0.5913, 0.4116,
        0.2247])
---------------------------------------
iter  0  stage  14  ep  0   adversary:  AdversaryModes.imitation_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
```

```
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0069, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.5460,
4.5460, 4.5460, 4.5460, 4.5460, 4.5460, 4.5460, 4.5460, 4.5460,
        4.5460, 4.5460, 4.5460, 4.5460, 4.5460, 4.5460, 3.9715, 3.4631, 3.0017,
        2.5741, 2.1709, 1.7856, 1.4135, 1.0510, 0.6958, 0.3459]) return=
135363.21866666665
probs of actions:  tensor([0.9988, 0.9987, 0.9984, 0.9989, 0.9992, 0.9992,
0.9989, 0.9989, 0.9989,
        0.9990, 0.9988, 0.9990, 0.9992, 0.9991, 0.9991, 0.9990, 0.9993, 0.9990,
        0.9997, 0.9995, 0.9996, 0.9997, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5381, 0.5432, 0.5419, 0.5422, 0.5421,
0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5746])
finalReturns:  tensor([1.4502, 1.4826, 1.4488, 1.3680, 1.2535, 1.1145, 0.9576,
0.7876, 0.6079,
        0.4209, 0.2287])
----------------------------------------
iter  0  stage  13  ep  0   adversary:  AdversaryModes.imitation_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0090, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.8889,
4.8889, 4.8889, 4.8889, 4.8889, 4.8889, 4.8889, 4.8889, 4.8889,
        4.8889, 4.8889, 4.8889, 4.8889, 4.8889, 4.3144, 3.8060, 3.3446, 2.9170,
        2.5138, 2.1285, 1.7564, 1.3940, 1.0387, 0.6888, 0.3429]) return=
135363.21866666665
probs of actions:  tensor([0.9988, 0.9987, 0.9984, 0.9989, 0.9992, 0.9992,
0.9989, 0.9989, 0.9989,
        0.9990, 0.9988, 0.9990, 0.9992, 0.9991, 0.9991, 0.9990, 0.9993, 0.9990,
        0.9997, 0.9995, 0.9996, 0.9997, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5381, 0.5432, 0.5419, 0.5422, 0.5421,
0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5746])
finalReturns:  tensor([1.6494, 1.6818, 1.6480, 1.5672, 1.4527, 1.3137, 1.1568,
0.9868, 0.8071,
        0.6202, 0.4279, 0.2316])
----------------------------------------
iter  0  stage  12  ep  0   adversary:  AdversaryModes.imitation_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0112, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.2297,
5.2297, 5.2297, 5.2297, 5.2297, 5.2297, 5.2297, 5.2297, 5.2297,
        5.2297, 5.2297, 5.2297, 5.2297, 4.6551, 4.1467, 3.6853, 3.2577, 2.8545,
```

```
        2.4693, 2.0971, 1.7347, 1.3794, 1.0295, 0.6836, 0.3407]) return=
135363.21866666665
probs of actions:  tensor([0.9988, 0.9987, 0.9984, 0.9989, 0.9992, 0.9992,
0.9989, 0.9989, 0.9989,
        0.9990, 0.9988, 0.9990, 0.9992, 0.9991, 0.9991, 0.9990, 0.9993, 0.9990,
        0.9997, 0.9995, 0.9996, 0.9997, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5381, 0.5432, 0.5419, 0.5422, 0.5421,
0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5746])
finalReturns:  tensor([1.8509, 1.8833, 1.8495, 1.7687, 1.6542, 1.5152, 1.3583,
1.1883, 1.0085,
        0.8216, 0.6293, 0.4331, 0.2339])
----------------------------------------
iter  0  stage  11  ep  0   adversary:  AdversaryModes.imitation_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0138, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.5687,
5.5687, 5.5687, 5.5687, 5.5687, 5.5687, 5.5687, 5.5687, 5.5687,
        5.5687, 5.5687, 5.5687, 4.9941, 4.4858, 4.0244, 3.5968, 3.1936, 2.8083,
        2.4362, 2.0737, 1.7185, 1.3686, 1.0227, 0.6798, 0.3390]) return=
135363.21866666665
probs of actions:  tensor([0.9988, 0.9987, 0.9984, 0.9989, 0.9992, 0.9992,
0.9989, 0.9989, 0.9989,
        0.9990, 0.9988, 0.9990, 0.9992, 0.9991, 0.9991, 0.9990, 0.9993, 0.9990,
        0.9997, 0.9995, 0.9996, 0.9997, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5381, 0.5432, 0.5419, 0.5422, 0.5421,
0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5746])
finalReturns:  tensor([2.0540, 2.0864, 2.0526, 1.9718, 1.8573, 1.7183, 1.5614,
1.3914, 1.2117,
        1.0247, 0.8325, 0.6362, 0.4370, 0.2355])
----------------------------------------
iter  0  stage  10  ep  52   adversary:  AdversaryModes.imitation_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0151, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.9065,
5.9065, 5.9065, 5.9065, 5.9065, 5.9065, 5.9065, 5.9065, 5.9065,
        5.9065, 5.9065, 5.3319, 4.8236, 4.3622, 3.9346, 3.5314, 3.1461, 2.7740,
        2.4115, 2.0563, 1.7064, 1.3605, 1.0176, 0.6768, 0.3378]) return=
135363.21866666665
probs of actions:  tensor([0.9988, 0.9988, 0.9985, 0.9989, 0.9992, 0.9992,
0.9990, 0.9990, 0.9990,
```

```
        0.9990, 0.9990, 0.9992, 0.9993, 0.9992, 0.9992, 0.9991, 0.9994, 0.9991,
        0.9997, 0.9995, 0.9996, 0.9997, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5381, 0.5432, 0.5419, 0.5422, 0.5421,
0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5746])
finalReturns:  tensor([2.2584, 2.2908, 2.2570, 2.1762, 2.0617, 1.9227, 1.7658,
1.5958, 1.4160,
        1.2291, 1.0368, 0.8406, 0.6413, 0.4399, 0.2368])
---------------------------------------
iter  0  stage  9  ep  0   adversary:  AdversaryModes.imitation_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0182, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.2434,
6.2434, 6.2434, 6.2434, 6.2434, 6.2434, 6.2434, 6.2434, 6.2434,
        6.2434, 5.6688, 5.1604, 4.6991, 4.2714, 3.8683, 3.4830, 3.1108, 2.7484,
        2.3932, 2.0433, 1.6974, 1.3544, 1.0137, 0.6747, 0.3369]) return=
135363.21866666665
probs of actions:  tensor([0.9988, 0.9988, 0.9985, 0.9989, 0.9992, 0.9992,
0.9990, 0.9990, 0.9990,
        0.9990, 0.9990, 0.9992, 0.9993, 0.9992, 0.9992, 0.9991, 0.9994, 0.9991,
        0.9997, 0.9995, 0.9996, 0.9997, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5381, 0.5432, 0.5419, 0.5422, 0.5421,
0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5746])
finalReturns:  tensor([2.4637, 2.4961, 2.4623, 2.3815, 2.2669, 2.1279, 1.9711,
1.8010, 1.6213,
        1.4344, 1.2421, 1.0459, 0.8466, 0.6452, 0.4421, 0.2377])
---------------------------------------
iter  0  stage  8  ep  6   adversary:  AdversaryModes.imitation_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0212, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.5795,
6.5795, 6.5795, 6.5795, 6.5795, 6.5795, 6.5795, 6.5795, 6.5795,
        6.0050, 5.4966, 5.0352, 4.6076, 4.2044, 3.8191, 3.4470, 3.0846, 2.7293,
        2.3794, 2.0335, 1.6906, 1.3499, 1.0108, 0.6730, 0.3362]) return=
135363.21866666665
probs of actions:  tensor([0.9988, 0.9988, 0.9985, 0.9990, 0.9993, 0.9992,
0.9990, 0.9990, 0.9990,
        0.9991, 0.9990, 0.9992, 0.9993, 0.9992, 0.9993, 0.9991, 0.9994, 0.9991,
        0.9997, 0.9995, 0.9996, 0.9997, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5381, 0.5432, 0.5419, 0.5422, 0.5421,
```

```
0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5746])
finalReturns:  tensor([2.6696, 2.7020, 2.6682, 2.5875, 2.4729, 2.3339, 2.1771,
2.0070, 1.8273,
        1.6404, 1.4481, 1.2519, 1.0526, 0.8512, 0.6480, 0.4437, 0.2384])
----------------------------------------
iter  0  stage  7  ep  11   adversary:  AdversaryModes.imitation_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0241, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.9152,
6.9152, 6.9152, 6.9152, 6.9152, 6.9152, 6.9152, 6.9152, 6.3406,
        5.8323, 5.3709, 4.9432, 4.5401, 4.1548, 3.7826, 3.4202, 3.0650, 2.7151,
        2.3692, 2.0262, 1.6855, 1.3465, 1.0087, 0.6718, 0.3356]) return=
135363.21866666665
probs of actions:  tensor([0.9989, 0.9988, 0.9986, 0.9990, 0.9993, 0.9992,
0.9990, 0.9990, 0.9990,
        0.9991, 0.9991, 0.9992, 0.9993, 0.9992, 0.9993, 0.9992, 0.9994, 0.9991,
        0.9997, 0.9995, 0.9996, 0.9997, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5381, 0.5432, 0.5419, 0.5422, 0.5421,
0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5746])
finalReturns:  tensor([2.8762, 2.9086, 2.8748, 2.7940, 2.6794, 2.5405, 2.3836,
2.2136, 2.0338,
        1.8469, 1.6546, 1.4584, 1.2591, 1.0577, 0.8546, 0.6502, 0.4449, 0.2389])
----------------------------------------
iter  0  stage  6  ep  0   adversary:  AdversaryModes.imitation_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0279, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.2504,
7.2504, 7.2504, 7.2504, 7.2504, 7.2504, 7.2504, 6.6759, 6.1675,
        5.7061, 5.2785, 4.8753, 4.4901, 4.1179, 3.7555, 3.4002, 3.0503, 2.7044,
        2.3615, 2.0208, 1.6818, 1.3440, 1.0071, 0.6709, 0.3353]) return=
135363.21866666665
probs of actions:  tensor([0.9989, 0.9988, 0.9986, 0.9990, 0.9993, 0.9992,
0.9990, 0.9990, 0.9990,
        0.9991, 0.9991, 0.9992, 0.9993, 0.9992, 0.9993, 0.9992, 0.9994, 0.9991,
        0.9997, 0.9995, 0.9996, 0.9997, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5381, 0.5432, 0.5419, 0.5422, 0.5421,
0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5746])
finalReturns:  tensor([3.0831, 3.1155, 3.0817, 3.0009, 2.8864, 2.7474, 2.5905,
```

```
2.4205, 2.2407,
        2.0538, 1.8615, 1.6653, 1.4660, 1.2646, 1.0615, 0.8571, 0.6518, 0.4458,
        0.2393])
---------------------------------------
iter  0  stage  5  ep  0   adversary:  AdversaryModes.imitation_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0311, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.5855,
7.5855, 7.5855, 7.5855, 7.5855, 7.5855, 7.0108, 6.5025, 6.0411,
        5.6135, 5.2103, 4.8250, 4.4529, 4.0904, 3.7352, 3.3853, 3.0394, 2.6965,
        2.3558, 2.0167, 1.6789, 1.3420, 1.0059, 0.6702, 0.3350]) return=
135363.21866666665
probs of actions:  tensor([0.9989, 0.9988, 0.9986, 0.9990, 0.9993, 0.9992,
0.9990, 0.9990, 0.9991,
        0.9991, 0.9991, 0.9992, 0.9993, 0.9992, 0.9993, 0.9992, 0.9994, 0.9991,
        0.9997, 0.9995, 0.9996, 0.9997, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5381, 0.5432, 0.5419, 0.5422, 0.5421,
0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5746])
finalReturns:  tensor([3.2903, 3.3227, 3.2889, 3.2081, 3.0936, 2.9546, 2.7977,
2.6277, 2.4479,
        2.2610, 2.0687, 1.8725, 1.6732, 1.4718, 1.2687, 1.0643, 0.8590, 0.6530,
        0.4465, 0.2396])
---------------------------------------
iter  0  stage  4  ep  0   adversary:  AdversaryModes.imitation_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0343, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.9200,
7.9200, 7.9200, 7.9200, 7.9200, 7.3456, 6.8372, 6.3759, 5.9482,
        5.5450, 5.1598, 4.7876, 4.4252, 4.0699, 3.7200, 3.3741, 3.0312, 2.6905,
        2.3515, 2.0137, 1.6768, 1.3406, 1.0050, 0.6697, 0.3347]) return=
135363.21866666665
probs of actions:  tensor([0.9989, 0.9988, 0.9986, 0.9990, 0.9993, 0.9992,
0.9990, 0.9990, 0.9991,
        0.9991, 0.9991, 0.9992, 0.9993, 0.9992, 0.9993, 0.9992, 0.9994, 0.9991,
        0.9997, 0.9995, 0.9996, 0.9997, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5381, 0.5432, 0.5419, 0.5422, 0.5421,
0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5746])
finalReturns:  tensor([3.4977, 3.5301, 3.4963, 3.4155, 3.3010, 3.1620, 3.0051,
2.8351, 2.6554,
        2.4684, 2.2762, 2.0799, 1.8807, 1.6792, 1.4761, 1.2717, 1.0664, 0.8604,
```

```
        0.6539, 0.4470, 0.2398])
---------------------------------------
iter  0  stage  3  ep  5   adversary:  AdversaryModes.imitation_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0380, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.2556,
8.2556, 8.2556, 8.2556, 7.6800, 7.1719, 6.7104, 6.2828, 5.8796,
        5.4943, 5.1222, 4.7598, 4.4045, 4.0546, 3.7087, 3.3658, 3.0251, 2.6860,
        2.3482, 2.0114, 1.6752, 1.3395, 1.0043, 0.6693, 0.3346]) return=
135363.21866666665
probs of actions:  tensor([0.9989, 0.9989, 0.9986, 0.9990, 0.9993, 0.9993,
0.9990, 0.9990, 0.9991,
        0.9991, 0.9991, 0.9992, 0.9994, 0.9993, 0.9993, 0.9992, 0.9994, 0.9991,
        0.9997, 0.9995, 0.9996, 0.9997, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5381, 0.5432, 0.5419, 0.5422, 0.5421,
0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5746])
finalReturns:  tensor([3.7053, 3.7377, 3.7039, 3.6231, 3.5086, 3.3696, 3.2127,
3.0427, 2.8629,
        2.6760, 2.4837, 2.2875, 2.0882, 1.8868, 1.6837, 1.4793, 1.2740, 1.0680,
        0.8615, 0.6546, 0.4474, 0.2400])
---------------------------------------
iter  0  stage  2  ep  223   adversary:  AdversaryModes.imitation_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0335, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.5859,
8.5859, 8.5859, 8.0154, 7.5061, 7.0449, 6.6172, 6.2141, 5.8288,
        5.4567, 5.0942, 4.7390, 4.3891, 4.0432, 3.7002, 3.3595, 3.0205, 2.6827,
        2.3458, 2.0097, 1.6740, 1.3387, 1.0038, 0.6690, 0.3345]) return=
135363.21866666665
probs of actions:  tensor([0.9990, 0.9990, 0.9990, 0.9993, 0.9995, 0.9995,
0.9992, 0.9992, 0.9993,
        0.9994, 0.9993, 0.9994, 0.9995, 0.9994, 0.9994, 0.9994, 0.9996, 0.9993,
        0.9998, 0.9996, 0.9997, 0.9997, 0.9999, 0.9999, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5381, 0.5432, 0.5419, 0.5422, 0.5421,
0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5746])
finalReturns:  tensor([3.9130, 3.9454, 3.9116, 3.8308, 3.7163, 3.5773, 3.4204,
3.2504, 3.0706,
        2.8837, 2.6914, 2.4952, 2.2960, 2.0945, 1.8914, 1.6870, 1.4817, 1.2757,
        1.0692, 0.8623, 0.6551, 0.4477, 0.2401])
---------------------------------------
```

```
iter  0  stage  1  ep  0   adversary:  AdversaryModes.imitation_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0382, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.9369,
8.9369, 8.3459, 7.8414, 7.3791, 6.9517, 6.5484, 6.1632, 5.7910,
        5.4286, 5.0733, 4.7234, 4.3775, 4.0346, 3.6939, 3.3549, 3.0171, 2.6802,
        2.3440, 2.0084, 1.6731, 1.3381, 1.0034, 0.6688, 0.3344]) return=
135363.21866666665
probs of actions:  tensor([0.9990, 0.9990, 0.9990, 0.9993, 0.9995, 0.9995,
0.9992, 0.9992, 0.9993,
        0.9994, 0.9993, 0.9994, 0.9995, 0.9994, 0.9994, 0.9994, 0.9996, 0.9993,
        0.9998, 0.9996, 0.9997, 0.9997, 0.9999, 0.9999, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5381, 0.5432, 0.5419, 0.5422, 0.5421,
0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5746])
finalReturns:  tensor([4.1206, 4.1530, 4.1195, 4.0386, 3.9241, 3.7851, 3.6282,
3.4582, 3.2784,
        3.0915, 2.8992, 2.7030, 2.5038, 2.3023, 2.0992, 1.8948, 1.6895, 1.4835,
        1.2770, 1.0701, 0.8629, 0.6555, 0.4479, 0.2402])
----------------------------------------
iter  0  stage  0  ep  0   adversary:  AdversaryModes.imitation_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0431, grad_fn=<NegBackward0>)   ,  base rewards= tensor([9.2070,
8.6958, 8.1720, 7.7142, 7.2857, 6.8828, 6.4974, 6.1253, 5.7629,
        5.4076, 5.0577, 4.7118, 4.3689, 4.0282, 3.6892, 3.3513, 3.0145, 2.6783,
        2.3427, 2.0074, 1.6724, 1.3377, 1.0031, 0.6687, 0.3343]) return=
135363.21866666665
probs of actions:  tensor([0.9990, 0.9990, 0.9990, 0.9993, 0.9995, 0.9995,
0.9992, 0.9992, 0.9993,
        0.9994, 0.9993, 0.9994, 0.9995, 0.9994, 0.9994, 0.9994, 0.9996, 0.9993,
        0.9998, 0.9996, 0.9997, 0.9997, 0.9999, 0.9999, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5381, 0.5432, 0.5419, 0.5422, 0.5421,
0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422,
        0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5422, 0.5746])
finalReturns:  tensor([4.3293, 4.3617, 4.3269, 4.2466, 4.1319, 3.9929, 3.8361,
3.6660, 3.4863,
        3.2994, 3.1071, 2.9109, 2.7116, 2.5102, 2.3071, 2.1027, 1.8974, 1.6914,
        1.4849, 1.2780, 1.0708, 0.8634, 0.6558, 0.4481, 0.2403])
0,[1e-05,1][1, 10000, 1, 1],1682701103 saved
[767881, 'tensor([0., 0., 0., 1., 0., 0., 0., 0., 0., 0., 0., 0.])',
135363.21866666665, 84740.85066666668, 0.043078526854515076, 1e-05, 1, 0,
```

```
'tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18,\n        18, 18, 18, 18, 18, 18,  0])', '[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.\n 1.]', '0,[1e-05,1][1, 10000, 1,
1],1682701103', 25, 50, 164632.33345850307, 193521.60851536895,
78111.91688437786, 135363.21866666665, 132442.06666666665, 102604.52812037412,
102604.52812037412, 120699.79187222247, 120677.7693195532, 90621.14382460734,
102604.52812037412, 120699.79187222247]
policy reset
----------------------------------------
iter  1  stage  24  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0,
        0])
loss=  tensor(-0., grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.5213,
0.5213, 0.5213, 0.5213, 0.5213, 0.5213, 0.5213, 0.5213, 0.5213,
        0.5213, 0.5213, 0.5213, 0.5213, 0.5213, 0.5213, 0.5213, 0.5213, 0.5213,
        0.5213, 0.5213, 0.5213, 0.5213, 0.5213, 0.5213, 0.5213]) return=
130240.65866666666
probs of actions:  tensor([0.9077, 0.8732, 0.8904, 0.8995, 0.8972, 0.8745,
0.8788, 0.9012, 0.8913,
        0.8767, 0.8811, 0.8843, 0.8777, 0.8816, 0.8965, 0.8712, 0.8952, 0.9004,
        0.8862, 0.9074, 0.8904, 0.8827, 0.8910, 0.8883, 0.9826],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5112, 0.5238, 0.5207, 0.5214, 0.5212, 0.5213, 0.5213,
0.5213, 0.5213,
        0.5213, 0.5213, 0.5213, 0.5213, 0.5213, 0.5213, 0.5213, 0.5213, 0.5213,
        0.5213, 0.5213, 0.5213, 0.5213, 0.5213, 0.5213, 0.5213])
finalReturns:  tensor([0.])
----------------------------------------
iter  1  stage  23  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([ 0,  0, 13,  0, 13,  0, 13,  0, 20,  3, 18,  3,  0, 18,  0,
15, 20,  0,
        0, 11,  0,  0, 17,  8,  0])
loss=  tensor(0.1189, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.0900,
1.0900, 1.0900, 1.0900, 1.0900, 1.0900, 1.0900, 1.0900, 1.0900,
        1.0900, 1.0900, 1.0900, 1.0900, 1.0900, 1.0900, 1.0900, 1.0900, 1.0900,
        1.0900, 1.0900, 1.0900, 1.0900, 1.0900, 1.0900, 0.5062]) return=
132730.16473578967
probs of actions:  tensor([0.5746, 0.4005, 0.0869, 0.5135, 0.0746, 0.3050,
0.0707, 0.4547, 0.0140,
        0.0303, 0.0911, 0.0262, 0.3726, 0.1150, 0.3055, 0.0646, 0.0140, 0.4273,
        0.4826, 0.0104, 0.4164, 0.3292, 0.0586, 0.0051, 0.9925],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5112, 0.5238, 0.5038, 0.5694, 0.4927, 0.5723, 0.4920,
0.5725, 0.4688,
        0.5984, 0.4809, 0.5895, 0.5154, 0.4904, 0.5879, 0.4828, 0.5411, 0.5805,
        0.5070, 0.5128, 0.5608, 0.5116, 0.4948, 0.5774, 0.5351])
finalReturns:  tensor([0.0225, 0.0289])
```

```
----------------------------------------
iter  1  stage  22  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 17, 18, 17, 17, 18, 18, 18, 23,
18, 17, 18,
        18, 18, 10, 17, 18, 18,  0])
loss=  tensor(0.1164, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.5474,
1.5474, 1.5474, 1.5474, 1.5474, 1.5474, 1.5474, 1.5474, 1.5474,
        1.5474, 1.5474, 1.5474, 1.5474, 1.5474, 1.5474, 1.5474, 1.5474, 1.5474,
        1.5474, 1.5474, 1.5474, 1.5474, 1.5474, 0.9691, 0.4616]) return=
135321.06489242485
probs of actions:  tensor([0.6817, 0.6622, 0.7054, 0.6411, 0.6841, 0.6765,
0.6575, 0.1339, 0.6236,
        0.1263, 0.1174, 0.6488, 0.6663, 0.7073, 0.0190, 0.6308, 0.1307, 0.6938,
        0.6575, 0.6494, 0.0023, 0.1165, 0.6519, 0.6349, 0.9946],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5381, 0.5432, 0.5419, 0.5422, 0.5421,
0.5457, 0.5384,
        0.5466, 0.5416, 0.5394, 0.5429, 0.5420, 0.5217, 0.5613, 0.5409, 0.5396,
        0.5428, 0.5420, 0.5646, 0.5157, 0.5460, 0.5412, 0.5748])
finalReturns:  tensor([0.1145, 0.1469, 0.1132])
----------------------------------------
iter  1  stage  21  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([17, 23, 15, 18, 20, 23, 18, 20, 18, 20, 17, 20, 18, 18, 20,
19, 25, 18,
        23, 18, 20, 18, 20, 20,  0])
loss=  tensor(0.6976, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.9726,
1.9726, 1.9726, 1.9726, 1.9726, 1.9726, 1.9726, 1.9726, 1.9726,
        1.9726, 1.9726, 1.9726, 1.9726, 1.9726, 1.9726, 1.9726, 1.9726, 1.9726,
        1.9726, 1.9726, 1.9726, 1.9726, 1.3892, 0.8829, 0.4210]) return=
135023.68474879788
probs of actions:  tensor([0.0946, 0.0818, 0.0169, 0.4284, 0.2996, 0.0770,
0.3800, 0.2651, 0.3633,
        0.2819, 0.1024, 0.2574, 0.4822, 0.4497, 0.3131, 0.0237, 0.0440, 0.4993,
        0.0806, 0.4510, 0.2391, 0.3161, 0.4021, 0.4613, 0.9969],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4823, 0.5342, 0.5680, 0.5269, 0.5384, 0.5283, 0.5596,
0.5302, 0.5509,
        0.5324, 0.5538, 0.5288, 0.5512, 0.5399, 0.5351, 0.5459, 0.5140, 0.5685,
        0.5152, 0.5629, 0.5294, 0.5511, 0.5323, 0.5427, 0.5801])
finalReturns:  tensor([0.2336, 0.2660, 0.2400, 0.1591])
----------------------------------------
iter  1  stage  20  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([25, 25, 23, 20, 23, 25, 18, 23, 18, 23, 18, 23, 23, 24, 17,
18, 23, 25,
        25, 23, 19, 25, 25, 20,  0])
loss=  tensor(3.1660, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.3248,
2.3248, 2.3248, 2.3248, 2.3248, 2.3248, 2.3248, 2.3248, 2.3248,
        2.3248, 2.3248, 2.3248, 2.3248, 2.3248, 2.3248, 2.3248, 2.3248, 2.3248,
```

```
        2.3248, 2.3248, 2.3248, 1.7364, 1.2313, 0.7725, 0.3692]) return=
134249.1143811909
probs of actions:  tensor([0.1810, 0.2000, 0.2224, 0.3108, 0.2469, 0.2092,
0.1259, 0.2311, 0.1225,
        0.2222, 0.2016, 0.2075, 0.2114, 0.0136, 0.0336, 0.1335, 0.2390, 0.2126,
        0.2172, 0.2367, 0.0198, 0.2455, 0.2390, 0.4338, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5557, 0.5376, 0.5497, 0.5255, 0.5302, 0.5644,
0.5162, 0.5627,
        0.5166, 0.5626, 0.5166, 0.5421, 0.5309, 0.5651, 0.5336, 0.5238, 0.5306,
        0.5342, 0.5429, 0.5522, 0.5124, 0.5388, 0.5547, 0.5772])
finalReturns:  tensor([0.4105, 0.4466, 0.4394, 0.3593, 0.2079])
----------------------------------------
iter  1  stage  19  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([25, 18, 25, 25, 25, 25, 25, 23, 25, 25, 25, 25, 26, 25, 25,
20, 25, 25,
        25, 25, 25, 25, 25, 20,  0])
loss=  tensor(1.5108, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.6317,
2.6317, 2.6317, 2.6317, 2.6317, 2.6317, 2.6317, 2.6317, 2.6317,
        2.6317, 2.6317, 2.6317, 2.6317, 2.6317, 2.6317, 2.6317, 2.6317, 2.6317,
        2.6317, 2.6317, 2.0354, 1.5321, 1.0930, 0.6993, 0.3379]) return=
133536.32186988593
probs of actions:  tensor([0.7044, 0.0200, 0.7188, 0.7269, 0.7118, 0.7290,
0.7193, 0.1459, 0.7147,
        0.7104, 0.7352, 0.7272, 0.0040, 0.7251, 0.7320, 0.0698, 0.7153, 0.7446,
        0.7193, 0.7999, 0.8080, 0.7683, 0.7679, 0.1276, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5858, 0.5014, 0.5416, 0.5315, 0.5340, 0.5334,
0.5431, 0.5258,
        0.5354, 0.5330, 0.5336, 0.5284, 0.5374, 0.5325, 0.5562, 0.5143, 0.5383,
        0.5323, 0.5338, 0.5334, 0.5335, 0.5335, 0.5560, 0.5768])
finalReturns:  tensor([0.6353, 0.6978, 0.6677, 0.5733, 0.4336, 0.2389])
----------------------------------------
iter  1  stage  18  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([25, 25, 23, 25, 25, 23, 25, 25, 25, 25, 25, 25, 25, 25, 25,
23, 23, 25,
        25, 25, 25, 25, 25, 25,  0])
loss=  tensor(0.3647, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.9535,
2.9535, 2.9535, 2.9535, 2.9535, 2.9535, 2.9535, 2.9535, 2.9535,
        2.9535, 2.9535, 2.9535, 2.9535, 2.9535, 2.9535, 2.9535, 2.9535, 2.9535,
        2.9535, 2.3561, 1.8530, 1.4139, 1.0202, 0.6589, 0.3209]) return=
133464.3841915384
probs of actions:  tensor([0.8935, 0.8888, 0.0682, 0.9030, 0.8959, 0.0677,
0.8989, 0.9129, 0.8946,
        0.8929, 0.9044, 0.9014, 0.9122, 0.9016, 0.9045, 0.0701, 0.0710, 0.9099,
        0.9199, 0.9389, 0.9494, 0.9240, 0.9188, 0.8524, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5557, 0.5376, 0.5272, 0.5351, 0.5427, 0.5259,
```

```
0.5354, 0.5330,
        0.5336, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5431, 0.5354, 0.5277,
        0.5349, 0.5331, 0.5336, 0.5335, 0.5335, 0.5335, 0.5960])
finalReturns:  tensor([0.8445, 0.9070, 0.8770, 0.7825, 0.6428, 0.4706, 0.2751])
---------------------------------------
iter  1  stage  17  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25,
25, 25, 25,
        25, 25, 25, 25, 25, 25,  0])
loss=  tensor(0.1831, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.2608,
3.2608, 3.2608, 3.2608, 3.2608, 3.2608, 3.2608, 3.2608, 3.2608,
        3.2608, 3.2608, 3.2608, 3.2608, 3.2608, 3.2608, 3.2608, 3.2608, 3.2608,
        2.6648, 2.1614, 1.7223, 1.3286, 0.9673, 0.6294, 0.3084]) return=
133326.59199999998
probs of actions:  tensor([0.9582, 0.9553, 0.9595, 0.9618, 0.9593, 0.9611,
0.9598, 0.9668, 0.9578,
        0.9583, 0.9635, 0.9620, 0.9660, 0.9618, 0.9629, 0.9578, 0.9580, 0.9654,
        0.9735, 0.9787, 0.9825, 0.9718, 0.9683, 0.9487, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5557, 0.5280, 0.5349, 0.5331, 0.5336, 0.5335,
0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5960])
finalReturns:  tensor([1.0696, 1.1321, 1.1020, 1.0076, 0.8678, 0.6957, 0.5001,
0.2875])
---------------------------------------
iter  1  stage  16  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25,
23, 23, 25,
        25, 25, 25, 25, 25, 25,  0])
loss=  tensor(4.6647, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.5711,
3.5711, 3.5711, 3.5711, 3.5711, 3.5711, 3.5711, 3.5711, 3.5711,
        3.5711, 3.5711, 3.5711, 3.5711, 3.5711, 3.5711, 3.5711, 3.5711, 2.9828,
        2.4776, 2.0323, 1.6354, 1.2714, 0.9317, 0.6094, 0.3000]) return=
133395.47270678813
probs of actions:  tensor([0.9645, 0.9623, 0.9662, 0.9679, 0.9659, 0.9673,
0.9660, 0.9719, 0.9648,
        0.9654, 0.9701, 0.9689, 0.9721, 0.9677, 0.9688, 0.0327, 0.0306, 0.9745,
        0.9786, 0.9827, 0.9879, 0.9764, 0.9713, 0.9507, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5557, 0.5280, 0.5349, 0.5331, 0.5336, 0.5335,
0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5431, 0.5354, 0.5277,
        0.5349, 0.5331, 0.5336, 0.5335, 0.5335, 0.5335, 0.5960])
finalReturns:  tensor([1.2901, 1.3430, 1.3204, 1.2308, 1.0946, 0.9250, 0.7313,
0.5201, 0.2960])
---------------------------------------
iter  1  stage  15  ep  99999   adversary:  AdversaryModes.imitation_132
```

```
  actions:  tensor([25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25,
25, 25, 25,
        25, 25, 25, 25, 25, 25,  0])
loss=  tensor(0.2434, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.8525,
3.8525, 3.8525, 3.8525, 3.8525, 3.8525, 3.8525, 3.8525, 3.8525,
        3.8525, 3.8525, 3.8525, 3.8525, 3.8525, 3.8525, 3.8525, 3.2565, 2.7531,
        2.3140, 1.9203, 1.5590, 1.2210, 0.9001, 0.5917, 0.2924]) return=
133326.59199999998
probs of actions:  tensor([0.9686, 0.9667, 0.9713, 0.9722, 0.9703, 0.9719,
0.9702, 0.9756, 0.9696,
        0.9702, 0.9744, 0.9733, 0.9758, 0.9720, 0.9730, 0.9731, 0.9757, 0.9777,
        0.9782, 0.9864, 0.9914, 0.9763, 0.9761, 0.9581, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5557, 0.5280, 0.5349, 0.5331, 0.5336, 0.5335,
0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5960])
finalReturns:  tensor([1.5449, 1.6074, 1.5773, 1.4829, 1.3431, 1.1710, 0.9754,
0.7628, 0.5378,
        0.3036])
----------------------------------------
iter  1  stage  14  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([23, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25,
25, 25, 25,
        25, 25, 25, 25, 25, 25,  0])
loss=  tensor(0.1918, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.1398,
4.1398, 4.1398, 4.1398, 4.1398, 4.1398, 4.1398, 4.1398, 4.1398,
        4.1398, 4.1398, 4.1398, 4.1398, 4.1398, 4.1398, 3.5438, 3.0405, 2.6014,
        2.2077, 1.8463, 1.5084, 1.1875, 0.8791, 0.5798, 0.2874]) return=
133359.57866666667
probs of actions:  tensor([0.0204, 0.9779, 0.9806, 0.9816, 0.9804, 0.9811,
0.9801, 0.9839, 0.9800,
        0.9800, 0.9831, 0.9827, 0.9841, 0.9813, 0.9832, 0.9834, 0.9820, 0.9878,
        0.9881, 0.9937, 0.9951, 0.9868, 0.9850, 0.9698, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4583, 0.5479, 0.5299, 0.5344, 0.5333, 0.5335, 0.5335,
0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5960])
finalReturns:  tensor([1.7910, 1.8535, 1.8234, 1.7290, 1.5892, 1.4171, 1.2215,
1.0089, 0.7839,
        0.5497, 0.3086])
----------------------------------------
iter  1  stage  13  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25,
25, 25, 25,
        25, 25, 25, 25, 25, 25,  0])
loss=  tensor(0.1643, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.4235,
```

```
4.4235, 4.4235, 4.4235, 4.4235, 4.4235, 4.4235, 4.4235, 4.4235,
        4.4235, 4.4235, 4.4235, 4.4235, 4.4235, 3.8275, 3.3241, 2.8850, 2.4913,
        2.1300, 1.7920, 1.4711, 1.1627, 0.8634, 0.5710, 0.2836]) return=
133326.59199999998
probs of actions:  tensor([0.9850, 0.9838, 0.9855, 0.9863, 0.9856, 0.9861,
0.9855, 0.9883, 0.9852,
        0.9854, 0.9877, 0.9874, 0.9883, 0.9897, 0.9887, 0.9869, 0.9845, 0.9928,
        0.9936, 0.9961, 0.9976, 0.9919, 0.9882, 0.9812, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5557, 0.5280, 0.5349, 0.5331, 0.5336, 0.5335,
0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5960])
finalReturns:  tensor([2.0409, 2.1034, 2.0733, 1.9789, 1.8391, 1.6669, 1.4714,
1.2588, 1.0338,
        0.7995, 0.5585, 0.3124])
----------------------------------------
iter  1  stage  12  ep  96178   adversary:  AdversaryModes.imitation_132
  actions:  tensor([25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25,
25, 25, 25,
        25, 25, 25, 25, 25, 25,  0])
loss=  tensor(0.0253, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.7043,
4.7043, 4.7043, 4.7043, 4.7043, 4.7043, 4.7043, 4.7043, 4.7043,
        4.7043, 4.7043, 4.7043, 4.7043, 4.1083, 3.6049, 3.1658, 2.7721, 2.4108,
        2.0728, 1.7519, 1.4435, 1.1442, 0.8518, 0.5644, 0.2808]) return=
133326.59199999998
probs of actions:  tensor([0.9974, 0.9971, 0.9973, 0.9975, 0.9975, 0.9975,
0.9974, 0.9980, 0.9973,
        0.9974, 0.9978, 0.9977, 0.9990, 0.9988, 0.9986, 0.9978, 0.9972, 0.9993,
        0.9994, 0.9997, 1.0000, 0.9991, 0.9988, 0.9974, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5557, 0.5280, 0.5349, 0.5331, 0.5336, 0.5335,
0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5960])
finalReturns:  tensor([2.2935, 2.3560, 2.3259, 2.2315, 2.0918, 1.9196, 1.7240,
1.5115, 1.2864,
        1.0522, 0.8111, 0.5650, 0.3152])
----------------------------------------
iter  1  stage  11  ep  85709   adversary:  AdversaryModes.imitation_132
  actions:  tensor([25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25,
25, 25, 25,
        25, 25, 25, 25, 25, 25,  0])
loss=  tensor(0.0115, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.9830,
4.9830, 4.9830, 4.9830, 4.9830, 4.9830, 4.9830, 4.9830, 4.9830,
        4.9830, 4.9830, 4.9830, 4.3870, 3.8836, 3.4445, 3.0508, 2.6895, 2.3516,
        2.0306, 1.7222, 1.4230, 1.1305, 0.8432, 0.5595, 0.2787]) return=
133326.59199999998
```

```
probs of actions:  tensor([0.9985, 0.9983, 0.9985, 0.9986, 0.9986, 0.9986,
0.9986, 0.9989, 0.9985,
        0.9985, 0.9988, 0.9990, 0.9996, 0.9997, 0.9994, 0.9992, 0.9990, 1.0000,
        0.9998, 1.0000, 1.0000, 0.9997, 0.9998, 0.9988, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5557, 0.5280, 0.5349, 0.5331, 0.5336, 0.5335,
0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5960])
finalReturns:  tensor([2.5483, 2.6108, 2.5807, 2.4863, 2.3465, 2.1744, 1.9788,
1.7662, 1.5412,
        1.3070, 1.0659, 0.8198, 0.5699, 0.3173])
----------------------------------------
iter  1  stage  10  ep  132   adversary:  AdversaryModes.imitation_132
  actions:  tensor([25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25,
25, 25, 25,
        25, 25, 25, 25, 25, 25,  0])
loss=  tensor(0.0145, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.2601,
5.2601, 5.2601, 5.2601, 5.2601, 5.2601, 5.2601, 5.2601, 5.2601,
        5.2601, 5.2601, 4.6642, 4.1608, 3.7217, 3.3280, 2.9667, 2.6287, 2.3078,
        1.9994, 1.7001, 1.4077, 1.1203, 0.8367, 0.5559, 0.2772]) return=
133326.59199999998
probs of actions:  tensor([0.9986, 0.9984, 0.9986, 0.9987, 0.9987, 0.9987,
0.9986, 0.9989, 0.9985,
        0.9986, 0.9990, 0.9990, 0.9996, 0.9997, 0.9995, 0.9992, 0.9990, 1.0000,
        0.9998, 1.0000, 1.0000, 0.9998, 0.9998, 0.9988, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5557, 0.5280, 0.5349, 0.5331, 0.5336, 0.5335,
0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5960])
finalReturns:  tensor([2.8046, 2.8671, 2.8370, 2.7426, 2.6028, 2.4307, 2.2351,
2.0226, 1.7975,
        1.5633, 1.3222, 1.0761, 0.8263, 0.5736, 0.3188])
----------------------------------------
iter  1  stage  9  ep  2917   adversary:  AdversaryModes.imitation_132
  actions:  tensor([25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25,
25, 25, 25,
        25, 25, 25, 25, 25, 25,  0])
loss=  tensor(0.0123, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.5361,
5.5361, 5.5361, 5.5361, 5.5361, 5.5361, 5.5361, 5.5361, 5.5361,
        5.5361, 4.9401, 4.4367, 3.9977, 3.6040, 3.2426, 2.9047, 2.5838, 2.2753,
        1.9761, 1.6837, 1.3963, 1.1127, 0.8319, 0.5531, 0.2760]) return=
133326.59199999998
probs of actions:  tensor([0.9990, 0.9988, 0.9990, 0.9990, 0.9990, 0.9990,
0.9990, 0.9992, 0.9989,
        0.9990, 0.9997, 0.9994, 0.9997, 0.9998, 0.9996, 0.9995, 0.9993, 1.0000,
        0.9999, 1.0000, 1.0000, 0.9998, 0.9999, 0.9991, 1.0000],
```

```
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5557, 0.5280, 0.5349, 0.5331, 0.5336, 0.5335,
0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5960])
finalReturns:  tensor([3.0621, 3.1246, 3.0945, 3.0001, 2.8603, 2.6882, 2.4926,
2.2801, 2.0550,
        1.8208, 1.5797, 1.3336, 1.0838, 0.8311, 0.5763, 0.3200])
----------------------------------------
iter  1  stage  8  ep  76   adversary:  AdversaryModes.imitation_132
  actions:  tensor([25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25,
25, 25, 25,
        25, 25, 25, 25, 25, 25,  0])
loss=  tensor(0.0164, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.8112,
5.8112, 5.8112, 5.8112, 5.8112, 5.8112, 5.8112, 5.8112, 5.8112,
        5.2152, 4.7119, 4.2728, 3.8791, 3.5177, 3.1798, 2.8589, 2.5504, 2.2512,
        1.9588, 1.6714, 1.3878, 1.1070, 0.8282, 0.5511, 0.2751]) return=
133326.59199999998
probs of actions:  tensor([0.9990, 0.9988, 0.9989, 0.9990, 0.9990, 0.9990,
0.9990, 0.9992, 0.9990,
        0.9990, 0.9997, 0.9994, 0.9997, 0.9998, 0.9996, 0.9995, 0.9993, 1.0000,
        0.9999, 1.0000, 1.0000, 0.9998, 0.9999, 0.9991, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5557, 0.5280, 0.5349, 0.5331, 0.5336, 0.5335,
0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5960])
finalReturns:  tensor([3.3205, 3.3830, 3.3529, 3.2585, 3.1187, 2.9466, 2.7510,
2.5385, 2.3134,
        2.0792, 1.8381, 1.5920, 1.3421, 1.0895, 0.8347, 0.5784, 0.3209])
----------------------------------------
iter  1  stage  7  ep  0   adversary:  AdversaryModes.imitation_132
  actions:  tensor([25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25,
25, 25, 25,
        25, 25, 25, 25, 25, 25,  0])
loss=  tensor(0.0199, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.0857,
6.0857, 6.0857, 6.0857, 6.0857, 6.0857, 6.0857, 6.0857, 5.4897,
        4.9863, 4.5472, 4.1535, 3.7922, 3.4543, 3.1333, 2.8249, 2.5257, 2.2332,
        1.9458, 1.6622, 1.3814, 1.1027, 0.8255, 0.5496, 0.2744]) return=
133326.59199999998
probs of actions:  tensor([0.9990, 0.9988, 0.9989, 0.9990, 0.9990, 0.9990,
0.9990, 0.9992, 0.9990,
        0.9990, 0.9997, 0.9994, 0.9997, 0.9998, 0.9996, 0.9995, 0.9993, 1.0000,
        0.9999, 1.0000, 1.0000, 0.9998, 0.9999, 0.9991, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5557, 0.5280, 0.5349, 0.5331, 0.5336, 0.5335,
0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335,
```

```
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5960])
finalReturns:  tensor([3.5795, 3.6420, 3.6119, 3.5175, 3.3778, 3.2056, 3.0101,
2.7975, 2.5724,
        2.3382, 2.0971, 1.8510, 1.6012, 1.3485, 1.0937, 0.8374, 0.5799, 0.3215])
----------------------------------------
iter  1  stage  6  ep  13   adversary:  AdversaryModes.imitation_132
  actions:  tensor([25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25,
25, 25, 25,
        25, 25, 25, 25, 25, 25,  0])
loss=  tensor(0.0244, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.3596,
6.3596, 6.3596, 6.3596, 6.3596, 6.3596, 6.3596, 5.7637, 5.2603,
        4.8212, 4.4275, 4.0662, 3.7282, 3.4073, 3.0989, 2.7996, 2.5072, 2.2198,
        1.9362, 1.6554, 1.3767, 1.0995, 0.8235, 0.5484, 0.2740]) return=
133326.59199999998
probs of actions:  tensor([0.9990, 0.9988, 0.9989, 0.9990, 0.9990, 0.9990,
0.9990, 0.9992, 0.9990,
        0.9990, 0.9997, 0.9994, 0.9997, 0.9998, 0.9996, 0.9995, 0.9993, 1.0000,
        0.9999, 1.0000, 1.0000, 0.9998, 0.9999, 0.9991, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5557, 0.5280, 0.5349, 0.5331, 0.5336, 0.5335,
0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5960])
finalReturns:  tensor([3.8391, 3.9016, 3.8715, 3.7771, 3.6373, 3.4651, 3.2696,
3.0570, 2.8320,
        2.5977, 2.3567, 2.1106, 1.8607, 1.6080, 1.3533, 1.0969, 0.8394, 0.5811,
        0.3220])
----------------------------------------
iter  1  stage  5  ep  7131   adversary:  AdversaryModes.imitation_132
  actions:  tensor([25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25,
25, 25, 25,
        25, 25, 25, 25, 25, 25,  0])
loss=  tensor(0.0227, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.6333,
6.6333, 6.6333, 6.6333, 6.6333, 6.6333, 6.0372, 5.5339, 5.0948,
        4.7011, 4.3397, 4.0018, 3.6809, 3.3724, 3.0732, 2.7808, 2.4934, 2.2098,
        1.9290, 1.6502, 1.3731, 1.0971, 0.8220, 0.5475, 0.2736]) return=
133326.59199999998
probs of actions:  tensor([0.9990, 0.9989, 0.9990, 0.9990, 0.9991, 0.9990,
0.9997, 0.9994, 0.9993,
        0.9991, 0.9998, 0.9994, 0.9998, 0.9999, 0.9996, 0.9997, 0.9993, 1.0000,
        0.9999, 1.0000, 1.0000, 0.9999, 1.0000, 0.9992, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5557, 0.5280, 0.5349, 0.5331, 0.5336, 0.5335,
0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5960])
finalReturns:  tensor([4.0990, 4.1615, 4.1314, 4.0369, 3.8972, 3.7250, 3.5295,
3.3169, 3.0919,
```

```
        2.8576, 2.6166, 2.3705, 2.1206, 1.8679, 1.6132, 1.3568, 1.0993, 0.8410,
        0.5819, 0.3224])
----------------------------------------
iter  1  stage  4  ep  0   adversary:  AdversaryModes.imitation_132
  actions:  tensor([25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25,
25, 25, 25,
        25, 25, 25, 25, 25, 25,  0])
loss=  tensor(0.0274, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.9063,
6.9063, 6.9063, 6.9063, 6.9063, 6.3106, 5.8071, 5.3681, 4.9744,
        4.6131, 4.2751, 3.9542, 3.6458, 3.3465, 3.0541, 2.7667, 2.4831, 2.2023,
        1.9236, 1.6464, 1.3704, 1.0953, 0.8209, 0.5469, 0.2733]) return=
133326.59199999998
probs of actions:  tensor([0.9990, 0.9989, 0.9990, 0.9990, 0.9991, 0.9990,
0.9997, 0.9994, 0.9993,
        0.9991, 0.9998, 0.9994, 0.9998, 0.9999, 0.9996, 0.9997, 0.9993, 1.0000,
        0.9999, 1.0000, 1.0000, 0.9999, 1.0000, 0.9992, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5557, 0.5280, 0.5349, 0.5331, 0.5336, 0.5335,
0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5960])
finalReturns:  tensor([4.3591, 4.4216, 4.3915, 4.2971, 4.1574, 3.9852, 3.7896,
3.5771, 3.3520,
        3.1178, 2.8767, 2.6306, 2.3808, 2.1281, 1.8733, 1.6170, 1.3595, 1.1011,
        0.8421, 0.5826, 0.3227])
----------------------------------------
iter  1  stage  3  ep  0   adversary:  AdversaryModes.imitation_132
  actions:  tensor([25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25,
25, 25, 25,
        25, 25, 25, 25, 25, 25,  0])
loss=  tensor(0.0327, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.1808,
7.1808, 7.1808, 7.1808, 6.5834, 6.0803, 5.6412, 5.2475, 4.8862,
        4.5482, 4.2273, 3.9189, 3.6196, 3.3272, 3.0398, 2.7562, 2.4754, 2.1967,
        1.9195, 1.6435, 1.3684, 1.0940, 0.8200, 0.5464, 0.2731]) return=
133326.59199999998
probs of actions:  tensor([0.9990, 0.9989, 0.9990, 0.9990, 0.9991, 0.9990,
0.9997, 0.9994, 0.9993,
        0.9991, 0.9998, 0.9994, 0.9998, 0.9999, 0.9996, 0.9997, 0.9993, 1.0000,
        0.9999, 1.0000, 1.0000, 0.9999, 1.0000, 0.9992, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5557, 0.5280, 0.5349, 0.5331, 0.5336, 0.5335,
0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5960])
finalReturns:  tensor([4.6195, 4.6820, 4.6519, 4.5575, 4.4177, 4.2456, 4.0500,
3.8375, 3.6124,
        3.3782, 3.1371, 2.8910, 2.6411, 2.3885, 2.1337, 1.8774, 1.6199, 1.3615,
        1.1025, 0.8429, 0.5831, 0.3229])
```

```
---------------------------------------
iter  1  stage  2  ep  9   adversary:  AdversaryModes.imitation_132
  actions:  tensor([25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25,
25, 25, 25,
        25, 25, 25, 25, 25, 25,  0])
loss=  tensor(0.0382, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.4481,
7.4481, 7.4481, 6.8576, 6.3530, 5.9142, 5.5204, 5.1591, 4.8212,
        4.5002, 4.1918, 3.8926, 3.6001, 3.3128, 3.0291, 2.7483, 2.4696, 2.1925,
        1.9165, 1.6414, 1.3669, 1.0930, 0.8194, 0.5461, 0.2729]) return=
133326.59199999998
probs of actions:  tensor([0.9990, 0.9989, 0.9990, 0.9991, 0.9991, 0.9990,
0.9997, 0.9994, 0.9993,
        0.9991, 0.9998, 0.9994, 0.9998, 0.9999, 0.9996, 0.9997, 0.9993, 1.0000,
        0.9999, 1.0000, 1.0000, 0.9999, 1.0000, 0.9992, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5557, 0.5280, 0.5349, 0.5331, 0.5336, 0.5335,
0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5960])
finalReturns:  tensor([4.8801, 4.9426, 4.9124, 4.8180, 4.6783, 4.5061, 4.3106,
4.0980, 3.8729,
        3.6387, 3.3977, 3.1515, 2.9017, 2.6490, 2.3942, 2.1379, 1.8804, 1.6220,
        1.3630, 1.1035, 0.8436, 0.5834, 0.3230])
---------------------------------------
iter  1  stage  1  ep  70   adversary:  AdversaryModes.imitation_132
  actions:  tensor([25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25,
25, 25, 25,
        25, 25, 25, 25, 25, 25,  0])
loss=  tensor(0.0420, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.7435,
7.7435, 7.1254, 6.6270, 6.1868, 5.7933, 5.4319, 5.0940, 4.7731,
        4.4646, 4.1654, 3.8730, 3.5856, 3.3020, 3.0212, 2.7424, 2.4653, 2.1893,
        1.9142, 1.6397, 1.3658, 1.0922, 0.8189, 0.5458, 0.2728]) return=
133326.59199999998
probs of actions:  tensor([0.9990, 0.9990, 0.9991, 0.9992, 0.9992, 0.9990,
0.9997, 0.9994, 0.9993,
        0.9991, 0.9998, 0.9994, 0.9998, 0.9999, 0.9996, 0.9997, 0.9993, 1.0000,
        0.9999, 1.0000, 1.0000, 0.9999, 1.0000, 0.9993, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5557, 0.5280, 0.5349, 0.5331, 0.5336, 0.5335,
0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5960])
finalReturns:  tensor([5.1404, 5.2029, 5.1732, 5.0786, 4.9389, 4.7668, 4.5712,
4.3586, 4.1336,
        3.8994, 3.6583, 3.4122, 3.1623, 2.9097, 2.6549, 2.3986, 2.1411, 1.8827,
        1.6237, 1.3641, 1.1042, 0.8441, 0.5837, 0.3232])
---------------------------------------
iter  1  stage  0  ep  0   adversary:  AdversaryModes.imitation_132
```

```
  actions:  tensor([25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25,
25, 25, 25,
        25, 25, 25, 25, 25, 25,  0])
loss=  tensor(0.0482, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.9300,
7.4188, 6.8950, 6.4606, 6.0658, 5.7047, 5.3667, 5.0458, 4.7374,
        4.4382, 4.1457, 3.8583, 3.5747, 3.2939, 3.0152, 2.7380, 2.4620, 2.1869,
        1.9125, 1.6385, 1.3649, 1.0916, 0.8185, 0.5456, 0.2727]) return=
133326.59199999998
probs of actions:  tensor([0.9990, 0.9990, 0.9991, 0.9992, 0.9992, 0.9990,
0.9997, 0.9994, 0.9993,
        0.9991, 0.9998, 0.9994, 0.9998, 0.9999, 0.9996, 0.9997, 0.9993, 1.0000,
        0.9999, 1.0000, 1.0000, 0.9999, 1.0000, 0.9993, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5557, 0.5280, 0.5349, 0.5331, 0.5336, 0.5335,
0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335,
        0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5335, 0.5960])
finalReturns:  tensor([5.4026, 5.4651, 5.4333, 5.3396, 5.1996, 5.0275, 4.8319,
4.6194, 4.3943,
        4.1601, 3.9190, 3.6729, 3.4231, 3.1704, 2.9156, 2.6593, 2.4018, 2.1434,
        1.8844, 1.6249, 1.3650, 1.1048, 0.8444, 0.5839, 0.3232])
0,[1e-05,1][1, 10000, 1, 1],1682740769 saved
[1632248, 'tensor([0., 0., 0., 1., 0., 0., 0., 0., 0., 0., 0., 0.])',
133326.59199999998, 77073.984, 0.0481695681810379, 1e-05, 1, 0, 'tensor([25, 25,
25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25,\n        25, 25,
25, 25, 25, 25,  0])', '[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
1. 1. 1. 1. 1. 1.\n 1.]', '0,[1e-05,1][1, 10000, 1, 1],1682740769', 25, 50,
171147.58350022632, 214073.728368001, 88670.91886744411, 133326.59199999998,
130350.63466666665, 79008.24733959991, 78991.42784922772, 95380.52812037412,
95380.52812037412, 102530.56714139864, 79008.24733959991, 95380.52812037412]
policy reset
----------------------------------------
iter  2  stage  24  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 9, 0, 0, 0,
0, 0, 0, 0,
        0])
loss=  tensor(-0., grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.5213,
0.5213, 0.5213, 0.5213, 0.5213, 0.5213, 0.5213, 0.5213, 0.5213,
        0.5213, 0.5213, 0.5213, 0.5213, 0.5213, 0.5213, 0.5213, 0.5213, 0.5213,
        0.5213, 0.5213, 0.5213, 0.5213, 0.5213, 0.5213, 0.5213]) return=
130508.83063279034
probs of actions:  tensor([0.8197, 0.8212, 0.8176, 0.8134, 0.8230, 0.8173,
0.8226, 0.1095, 0.8180,
        0.1128, 0.8094, 0.1104, 0.8382, 0.8347, 0.8563, 0.8227, 0.0014, 0.8433,
        0.7915, 0.8165, 0.8150, 0.8321, 0.8387, 0.7760, 0.9487],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5112, 0.5238, 0.5207, 0.5214, 0.5212, 0.5213, 0.5213,
0.5212, 0.5249,
```

```
        0.5203, 0.5251, 0.5202, 0.5251, 0.5203, 0.5215, 0.5212, 0.5132, 0.5543,
        0.5132, 0.5233, 0.5208, 0.5214, 0.5213, 0.5213, 0.5213])
finalReturns:  tensor([0.])
----------------------------------------
iter  2  stage  23  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([ 0,  0, 17,  0,  0,  2,  0,  0,  0,  0, 12, 17, 16,  3,  0,
0, 16,  0,
        17, 17,  0, 17, 17, 17,  0])
loss=  tensor(0.0142, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.0776,
1.0776, 1.0776, 1.0776, 1.0776, 1.0776, 1.0776, 1.0776, 1.0776, 1.0776,
        1.0776, 1.0776, 1.0776, 1.0776, 1.0776, 1.0776, 1.0776, 1.0776, 1.0776,
        1.0776, 1.0776, 1.0776, 1.0776, 1.0776, 1.0776, 0.5100]) return=
132640.40457803075
probs of actions:  tensor([4.3731e-01, 4.2664e-01, 3.0690e-01, 4.7486e-01,
6.2047e-01, 1.8283e-02,
        5.1649e-01, 4.6048e-01, 5.6530e-01, 5.5434e-01, 4.1340e-03, 2.0115e-01,
        1.6684e-02, 3.7736e-04, 4.3648e-01, 4.8998e-01, 1.9829e-02, 5.7272e-01,
        4.0393e-01, 3.1937e-01, 4.3911e-01, 3.3119e-01, 3.0537e-01, 6.6976e-01,
        9.8906e-01], grad_fn=<ExpBackward0>)
rewards:  tensor([0.5112, 0.5238, 0.4918, 0.5846, 0.5060, 0.5247, 0.5276,
0.5197, 0.5217,
        0.5212, 0.5069, 0.5366, 0.5475, 0.5665, 0.5208, 0.5214, 0.4957, 0.5807,
        0.4780, 0.5594, 0.5674, 0.4812, 0.5586, 0.5387, 0.5725])
finalReturns:  tensor([0.0336, 0.0625])
----------------------------------------
iter  2  stage  22  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([17, 17, 17, 17, 17, 17, 17, 17, 17, 26, 17, 17, 17, 17, 17,
17, 17, 17,
        17, 17, 17, 17, 17, 17,  0])
loss=  tensor(0.0192, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.5452,
1.5452, 1.5452, 1.5452, 1.5452, 1.5452, 1.5452, 1.5452, 1.5452, 1.5452,
        1.5452, 1.5452, 1.5452, 1.5452, 1.5452, 1.5452, 1.5452, 1.5452, 1.5452,
        1.5452, 1.5452, 1.5452, 1.5452, 1.5452, 1.5452, 0.9737, 0.4646]) return=
135360.55464815182
probs of actions:  tensor([0.9285, 0.9338, 0.9111, 0.9164, 0.8840, 0.9259,
0.8917, 0.9240, 0.9054,
        0.0012, 0.9157, 0.9091, 0.9244, 0.9328, 0.9384, 0.9183, 0.9283, 0.9249,
        0.9404, 0.9240, 0.9221, 0.9283, 0.9593, 0.9090, 0.9896],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4823, 0.5582, 0.5388, 0.5436, 0.5424, 0.5427, 0.5426,
0.5426, 0.5426,
        0.5039, 0.5772, 0.5342, 0.5448, 0.5421, 0.5428, 0.5426, 0.5426, 0.5426,
        0.5426, 0.5426, 0.5426, 0.5426, 0.5426, 0.5426, 0.5715])
finalReturns:  tensor([0.1116, 0.1405, 0.1069])
----------------------------------------
iter  2  stage  21  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([17, 24, 17, 17, 17, 17, 17, 22, 17, 17, 17, 17, 17, 17, 17,
17, 17, 22,
```

```
        17, 17, 22, 17, 17, 17,  0])
loss=  tensor(0.1288, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.9930,
1.9930, 1.9930, 1.9930, 1.9930, 1.9930, 1.9930, 1.9930, 1.9930,
        1.9930, 1.9930, 1.9930, 1.9930, 1.9930, 1.9930, 1.9930, 1.9930, 1.9930,
        1.9930, 1.9930, 1.9930, 1.9930, 1.4027, 0.8980, 0.4323]) return=
135269.9521378386
probs of actions:  tensor([0.8732, 0.0179, 0.8410, 0.8647, 0.8634, 0.8757,
0.8419, 0.0913, 0.8787,
        0.9062, 0.8654, 0.8757, 0.8808, 0.9079, 0.8892, 0.8849, 0.8757, 0.0644,
        0.8793, 0.8759, 0.0651, 0.8430, 0.8864, 0.7571, 0.9970],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4823, 0.5295, 0.5654, 0.5370, 0.5440, 0.5423, 0.5427,
0.5231, 0.5617,
        0.5379, 0.5438, 0.5423, 0.5427, 0.5426, 0.5426, 0.5426, 0.5426, 0.5231,
        0.5617, 0.5379, 0.5243, 0.5614, 0.5380, 0.5438, 0.5712])
finalReturns:  tensor([0.2214, 0.2503, 0.2170, 0.1389])
----------------------------------------
iter  2  stage  20  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([18, 22, 22, 22, 16, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22, 22, 22,
        22, 22, 22, 22, 17, 24,  0])
loss=  tensor(2.2023, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.3434,
2.3434, 2.3434, 2.3434, 2.3434, 2.3434, 2.3434, 2.3434, 2.3434,
        2.3434, 2.3434, 2.3434, 2.3434, 2.3434, 2.3434, 2.3434, 2.3434, 2.3434,
        2.3434, 2.3434, 2.3434, 1.7567, 1.2512, 0.8026, 0.3945]) return=
134596.2288854169
probs of actions:  tensor([0.0027, 0.8000, 0.8295, 0.7982, 0.0137, 0.7866,
0.8077, 0.8311, 0.7757,
        0.7676, 0.8022, 0.8268, 0.8164, 0.7579, 0.8175, 0.7962, 0.8125, 0.8182,
        0.8061, 0.8040, 0.8673, 0.8985, 0.0647, 0.0441, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5426, 0.5373, 0.5386, 0.5611, 0.5156, 0.5441,
0.5369, 0.5387,
        0.5383, 0.5384, 0.5384, 0.5384, 0.5384, 0.5384, 0.5384, 0.5384, 0.5384,
        0.5384, 0.5384, 0.5384, 0.5384, 0.5579, 0.5102, 0.5993])
finalReturns:  tensor([0.4006, 0.4490, 0.4161, 0.3068, 0.2047])
----------------------------------------
iter  2  stage  19  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([22, 22, 22, 22, 22, 17, 22, 22, 22, 22, 22, 24, 22, 22, 22,
22, 22, 22,
        22, 22, 22, 22, 22, 22,  0])
loss=  tensor(0.1875, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.6857,
2.6857, 2.6857, 2.6857, 2.6857, 2.6857, 2.6857, 2.6857, 2.6857,
        2.6857, 2.6857, 2.6857, 2.6857, 2.6857, 2.6857, 2.6857, 2.6857, 2.6857,
        2.6857, 2.6857, 2.6857, 2.0990, 1.5935, 1.1449, 0.7368, 0.3579]) return=
134488.70432871333
probs of actions:  tensor([0.9050, 0.9012, 0.9150, 0.8960, 0.8927, 0.0392,
0.9032, 0.9113, 0.8899,
```

```
        0.8879, 0.9010, 0.0538, 0.9051, 0.8822, 0.9117, 0.8996, 0.9057, 0.9102,
        0.9060, 0.9025, 0.9461, 0.9436, 0.9424, 0.9373, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5581, 0.5335, 0.5396, 0.5381, 0.5579, 0.5193,
0.5432, 0.5372,
        0.5387, 0.5383, 0.5292, 0.5460, 0.5364, 0.5388, 0.5382, 0.5384, 0.5383,
        0.5384, 0.5384, 0.5384, 0.5384, 0.5384, 0.5384, 0.5868])
finalReturns:  tensor([0.5928, 0.6412, 0.6084, 0.5186, 0.3883, 0.2289])
----------------------------------------
iter  2  stage  18  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 17,
22, 22, 22,
        22, 22, 22, 22, 22, 22,  0])
loss=  tensor(0.2667, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.0284,
3.0284, 3.0284, 3.0284, 3.0284, 3.0284, 3.0284, 3.0284, 3.0284,
        3.0284, 3.0284, 3.0284, 3.0284, 3.0284, 3.0284, 3.0284, 3.0284, 3.0284,
        3.0284, 2.4413, 1.9359, 1.4873, 1.0792, 0.7003, 0.3424]) return=
134519.1548127569
probs of actions:  tensor([0.9210, 0.9158, 0.9274, 0.9103, 0.9067, 0.9078,
0.9168, 0.9221, 0.9075,
        0.9069, 0.9156, 0.9256, 0.9163, 0.9014, 0.0181, 0.9141, 0.9187, 0.9227,
        0.9245, 0.9104, 0.9560, 0.9496, 0.9478, 0.9500, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5581, 0.5335, 0.5396, 0.5381, 0.5384, 0.5383,
0.5384, 0.5384,
        0.5384, 0.5384, 0.5384, 0.5384, 0.5384, 0.5579, 0.5194, 0.5432, 0.5372,
        0.5387, 0.5383, 0.5384, 0.5384, 0.5384, 0.5384, 0.5868])
finalReturns:  tensor([0.7887, 0.8371, 0.8043, 0.7145, 0.5842, 0.4249, 0.2443])
----------------------------------------
iter  2  stage  17  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([22, 24, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22, 22, 22,
        22, 22, 22, 22, 22, 22,  0])
loss=  tensor(0.2357, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.3592,
3.3592, 3.3592, 3.3592, 3.3592, 3.3592, 3.3592, 3.3592, 3.3592,
        3.3592, 3.3592, 3.3592, 3.3592, 3.3592, 3.3592, 3.3592, 3.3592, 3.3592,
        2.7724, 2.2669, 1.8184, 1.4103, 1.0313, 0.6735, 0.3310]) return=
134444.89466666666
probs of actions:  tensor([0.9528, 0.0494, 0.9549, 0.9443, 0.9409, 0.9427,
0.9482, 0.9500, 0.9457,
        0.9458, 0.9480, 0.9526, 0.9470, 0.9422, 0.9535, 0.9476, 0.9500, 0.9530,
        0.9664, 0.9401, 0.9671, 0.9680, 0.9685, 0.9721, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5489, 0.5411, 0.5377, 0.5385, 0.5383, 0.5384,
0.5384, 0.5384,
        0.5384, 0.5384, 0.5384, 0.5384, 0.5384, 0.5384, 0.5384, 0.5384, 0.5384,
        0.5384, 0.5384, 0.5384, 0.5384, 0.5384, 0.5384, 0.5868])
finalReturns:  tensor([0.9960, 1.0444, 1.0116, 0.9218, 0.7915, 0.6322, 0.4517,
```

```
0.2557])
---------------------------------------
iter  2  stage  16  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 24,
22, 22, 22,
        22, 22, 22, 22, 22, 22,  0])
loss=  tensor(0.2379, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.6803,
3.6803, 3.6803, 3.6803, 3.6803, 3.6803, 3.6803, 3.6803, 3.6803, 3.6803,
        3.6803, 3.6803, 3.6803, 3.6803, 3.6803, 3.6803, 3.6803, 3.6803, 3.0955,
        2.5895, 2.1410, 1.7329, 1.3540, 0.9961, 0.6537, 0.3227]) return=
134445.23460823056
probs of actions:  tensor([0.9648, 0.9603, 0.9664, 0.9577, 0.9548, 0.9563,
0.9609, 0.9619, 0.9596,
        0.9595, 0.9610, 0.9640, 0.9599, 0.9569, 0.0341, 0.9608, 0.9626, 0.9690,
        0.9761, 0.9543, 0.9755, 0.9771, 0.9764, 0.9821, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5581, 0.5335, 0.5396, 0.5381, 0.5384, 0.5383,
0.5384, 0.5384,
        0.5384, 0.5384, 0.5384, 0.5384, 0.5384, 0.5292, 0.5460, 0.5364, 0.5388,
        0.5382, 0.5384, 0.5383, 0.5384, 0.5384, 0.5384, 0.5868])
finalReturns:  tensor([1.2118, 1.2602, 1.2273, 1.1375, 1.0072, 0.8479, 0.6674,
0.4714, 0.2641])
---------------------------------------
iter  2  stage  15  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22, 22, 22,
        22, 22, 22, 22, 22, 22,  0])
loss=  tensor(0.2841, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.9983,
3.9983, 3.9983, 3.9983, 3.9983, 3.9983, 3.9983, 3.9983, 3.9983, 3.9983,
        3.9983, 3.9983, 3.9983, 3.9983, 3.9983, 3.9983, 3.9983, 3.4115, 2.9060,
        2.4574, 2.0494, 1.6704, 1.3125, 0.9701, 0.6391, 0.3164]) return=
134475.688
probs of actions:  tensor([0.9686, 0.9641, 0.9696, 0.9614, 0.9594, 0.9605,
0.9648, 0.9655, 0.9639,
        0.9637, 0.9650, 0.9671, 0.9636, 0.9615, 0.9688, 0.9677, 0.9654, 0.9739,
        0.9781, 0.9543, 0.9812, 0.9795, 0.9813, 0.9855, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5581, 0.5335, 0.5396, 0.5381, 0.5384, 0.5383,
0.5384, 0.5384,
        0.5384, 0.5384, 0.5384, 0.5384, 0.5384, 0.5384, 0.5384, 0.5384, 0.5384,
        0.5384, 0.5384, 0.5384, 0.5384, 0.5384, 0.5384, 0.5868])
finalReturns:  tensor([1.4337, 1.4821, 1.4492, 1.3595, 1.2292, 1.0698, 0.8893,
0.6933, 0.4860,
        0.2703])
---------------------------------------
iter  2  stage  14  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([22, 22, 22, 24, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22,
22, 22, 22,
```

```
        22, 22, 22, 22, 22, 22,  0])
loss=  tensor(0.3424, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.3101,
4.3101, 4.3101, 4.3101, 4.3101, 4.3101, 4.3101, 4.3101, 4.3101,
        4.3101, 4.3101, 4.3101, 4.3101, 4.3101, 4.3101, 3.7233, 3.2178, 2.7692,
        2.3612, 1.9822, 1.6243, 1.2819, 0.9509, 0.6282, 0.3118]) return=
134445.21341666667
probs of actions:  tensor([0.9685, 0.9639, 0.9698, 0.0383, 0.9595, 0.9606,
0.9645, 0.9653, 0.9641,
        0.9640, 0.9651, 0.9672, 0.9636, 0.9619, 0.9746, 0.9696, 0.9669, 0.9763,
        0.9773, 0.9544, 0.9801, 0.9832, 0.9823, 0.9843, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5581, 0.5335, 0.5304, 0.5457, 0.5365, 0.5388,
0.5382, 0.5384,
        0.5383, 0.5384, 0.5384, 0.5384, 0.5384, 0.5384, 0.5384, 0.5384, 0.5384,
        0.5384, 0.5384, 0.5384, 0.5384, 0.5384, 0.5384, 0.5868])
finalReturns:  tensor([1.6602, 1.7086, 1.6758, 1.5860, 1.4557, 1.2964, 1.1158,
0.9199, 0.7126,
        0.4969, 0.2750])
----------------------------------------
iter  2  stage  13  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([22, 22, 22, 22, 22, 24, 22, 24, 24, 22, 22, 22, 24, 22, 22,
22, 22, 22,
        24, 22, 24, 22, 22, 22,  0])
loss=  tensor(8.3404, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.5998,
4.5998, 4.5998, 4.5998, 4.5998, 4.5998, 4.5998, 4.5998, 4.5998,
        4.5998, 4.5998, 4.5998, 4.5998, 4.5998, 4.0054, 3.5016, 3.0526, 2.6447,
        2.2657, 1.9078, 1.5654, 1.2401, 0.9203, 0.6119, 0.3046]) return=
134292.63893419717
probs of actions:  tensor([0.9018, 0.8871, 0.9064, 0.8867, 0.8751, 0.1197,
0.8925, 0.1104, 0.1097,
        0.8910, 0.8925, 0.8960, 0.1160, 0.8909, 0.9092, 0.9009, 0.8661, 0.9170,
        0.0809, 0.8616, 0.0710, 0.9437, 0.9367, 0.9406, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5581, 0.5335, 0.5396, 0.5381, 0.5292, 0.5460,
0.5272, 0.5373,
        0.5440, 0.5370, 0.5387, 0.5291, 0.5461, 0.5364, 0.5388, 0.5382, 0.5384,
        0.5291, 0.5460, 0.5272, 0.5465, 0.5363, 0.5389, 0.5866])
finalReturns:  tensor([1.9089, 1.9573, 1.9246, 1.8347, 1.7045, 1.5451, 1.3738,
1.1702, 0.9682,
        0.7415, 0.5136, 0.2820])
----------------------------------------
iter  2  stage  12  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([22, 24, 24, 22, 22, 22, 24, 22, 22, 22, 22, 24, 24, 22, 22,
22, 22, 22,
        22, 22, 24, 24, 22, 24,  0])
loss=  tensor(12.3434, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([4.8923, 4.8923, 4.8923, 4.8923, 4.8923, 4.8923, 4.8923, 4.8923, 4.8923,
        4.8923, 4.8923, 4.8923, 4.8923, 4.2979, 3.7942, 3.3519, 2.9471, 2.5707,
```

```
        2.2148, 1.8738, 1.5437, 1.2218, 0.9060, 0.6001, 0.3004]) return=
134247.45269681388
probs of actions:  tensor([0.6391, 0.4036, 0.3466, 0.6153, 0.5722, 0.5878,
0.3735, 0.5996, 0.6154,
        0.6103, 0.6166, 0.3860, 0.4416, 0.5771, 0.5627, 0.6402, 0.5170, 0.6724,
        0.6441, 0.5184, 0.3052, 0.2713, 0.6895, 0.3034, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5489, 0.5319, 0.5453, 0.5366, 0.5388, 0.5290,
0.5461, 0.5364,
        0.5388, 0.5382, 0.5292, 0.5368, 0.5441, 0.5369, 0.5387, 0.5383, 0.5384,
        0.5384, 0.5384, 0.5292, 0.5368, 0.5441, 0.5277, 0.5948])
finalReturns:  tensor([2.1502, 2.2078, 2.1675, 2.0728, 1.9389, 1.7770, 1.5946,
1.3972, 1.1889,
        0.9816, 0.7607, 0.5224, 0.2944])
----------------------------------------
iter  2  stage  11  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([24, 24, 24, 24, 24, 22, 24, 22, 24, 24, 24, 24, 24, 24, 24,
24, 24, 24,
        24, 24, 22, 24, 24, 24,  0])
loss=  tensor(5.1378, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.0743,
5.0743, 5.0743, 5.0743, 5.0743, 5.0743, 5.0743, 5.0743, 5.0743,
        5.0743, 5.0743, 5.0743, 4.4812, 3.9772, 3.5349, 3.1365, 2.7693, 2.4248,
        2.0968, 1.7809, 1.4739, 1.1736, 0.8782, 0.5810, 0.2893]) return=
133847.94214272863
probs of actions:  tensor([0.8439, 0.8704, 0.8356, 0.8477, 0.8817, 0.1337,
0.8461, 0.1280, 0.8540,
        0.8596, 0.8526, 0.8777, 0.9175, 0.8723, 0.9118, 0.8651, 0.9045, 0.8369,
        0.8812, 0.9076, 0.1471, 0.8375, 0.8724, 0.8691, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4536, 0.5567, 0.5300, 0.5366, 0.5350, 0.5446, 0.5276,
0.5464, 0.5271,
        0.5373, 0.5348, 0.5354, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353,
        0.5353, 0.5353, 0.5445, 0.5276, 0.5372, 0.5348, 0.5930])
finalReturns:  tensor([2.4807, 2.5383, 2.5071, 2.4141, 2.2772, 2.1091, 1.9183,
1.7110, 1.4916,
        1.2632, 1.0191, 0.7869, 0.5468, 0.3037])
----------------------------------------
iter  2  stage  10  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([24, 24, 24, 24, 22, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24,
24, 24, 24,
        24, 24, 24, 24, 24, 24,  0])
loss=  tensor(0.4986, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.3491,
5.3491, 5.3491, 5.3491, 5.3491, 5.3491, 5.3491, 5.3491, 5.3491,
        5.3491, 5.3491, 4.7562, 4.2521, 3.8099, 3.4114, 3.0443, 2.6998, 2.3718,
        2.0559, 1.7489, 1.4486, 1.1532, 0.8614, 0.5724, 0.2855]) return=
133786.32760416673
probs of actions:  tensor([0.9711, 0.9771, 0.9691, 0.9704, 0.0206, 0.9748,
0.9702, 0.9777, 0.9723,
```

```
        0.9741, 0.9751, 0.9822, 0.9866, 0.9817, 0.9886, 0.9788, 0.9858, 0.9735,
        0.9832, 0.9838, 0.9789, 0.9763, 0.9846, 0.9826, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4536, 0.5567, 0.5300, 0.5366, 0.5442, 0.5277, 0.5372,
0.5348, 0.5354,
        0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353,
        0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5929])
finalReturns:  tensor([2.7380, 2.7956, 2.7644, 2.6713, 2.5345, 2.3663, 2.1755,
1.9682, 1.7488,
        1.5205, 1.2855, 1.0456, 0.8021, 0.5558, 0.3074])
---------------------------------------
iter  2  stage  9  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24,
24, 24, 24,
        24, 24, 24, 24, 24, 24,  0])
loss=  tensor(0.2965, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.6334,
5.6334, 5.6334, 5.6334, 5.6334, 5.6334, 5.6334, 5.6334, 5.6334, 5.6334,
        5.6334, 5.0405, 4.5364, 4.0942, 3.6957, 3.3286, 2.9841, 2.6561, 2.3402,
        2.0332, 1.7329, 1.4375, 1.1458, 0.8567, 0.5698, 0.2843]) return=
133755.66666666666
probs of actions:  tensor([0.9837, 0.9872, 0.9825, 0.9829, 0.9885, 0.9855,
0.9831, 0.9875, 0.9842,
        0.9881, 0.9885, 0.9909, 0.9935, 0.9897, 0.9942, 0.9897, 0.9917, 0.9867,
        0.9917, 0.9918, 0.9900, 0.9879, 0.9922, 0.9919, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4536, 0.5567, 0.5300, 0.5366, 0.5350, 0.5354, 0.5353,
0.5353, 0.5353,
        0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353,
        0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5929])
finalReturns:  tensor([2.9890, 3.0466, 3.0154, 2.9223, 2.7855, 2.6173, 2.4265,
2.2192, 1.9998,
        1.7715, 1.5365, 1.2966, 1.0530, 0.8068, 0.5584, 0.3086])
---------------------------------------
iter  2  stage  8  ep  99999   adversary:  AdversaryModes.imitation_132
  actions:  tensor([24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24,
24, 24, 24,
        24, 24, 24, 24, 24, 24,  0])
loss=  tensor(0.0961, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.9169,
5.9169, 5.9169, 5.9169, 5.9169, 5.9169, 5.9169, 5.9169, 5.9169, 5.9169,
        5.3240, 4.8199, 4.3777, 3.9792, 3.6120, 3.2675, 2.9395, 2.6236, 2.3167,
        2.0164, 1.7210, 1.4292, 1.1402, 0.8532, 0.5678, 0.2835]) return=
133755.66666666666
probs of actions:  tensor([0.9944, 0.9956, 0.9939, 0.9940, 0.9961, 0.9949,
0.9941, 0.9957, 0.9959,
        0.9963, 0.9971, 0.9972, 0.9983, 0.9974, 0.9986, 0.9978, 0.9971, 0.9966,
        0.9973, 0.9980, 0.9976, 0.9971, 0.9983, 0.9984, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4536, 0.5567, 0.5300, 0.5366, 0.5350, 0.5354, 0.5353,
```

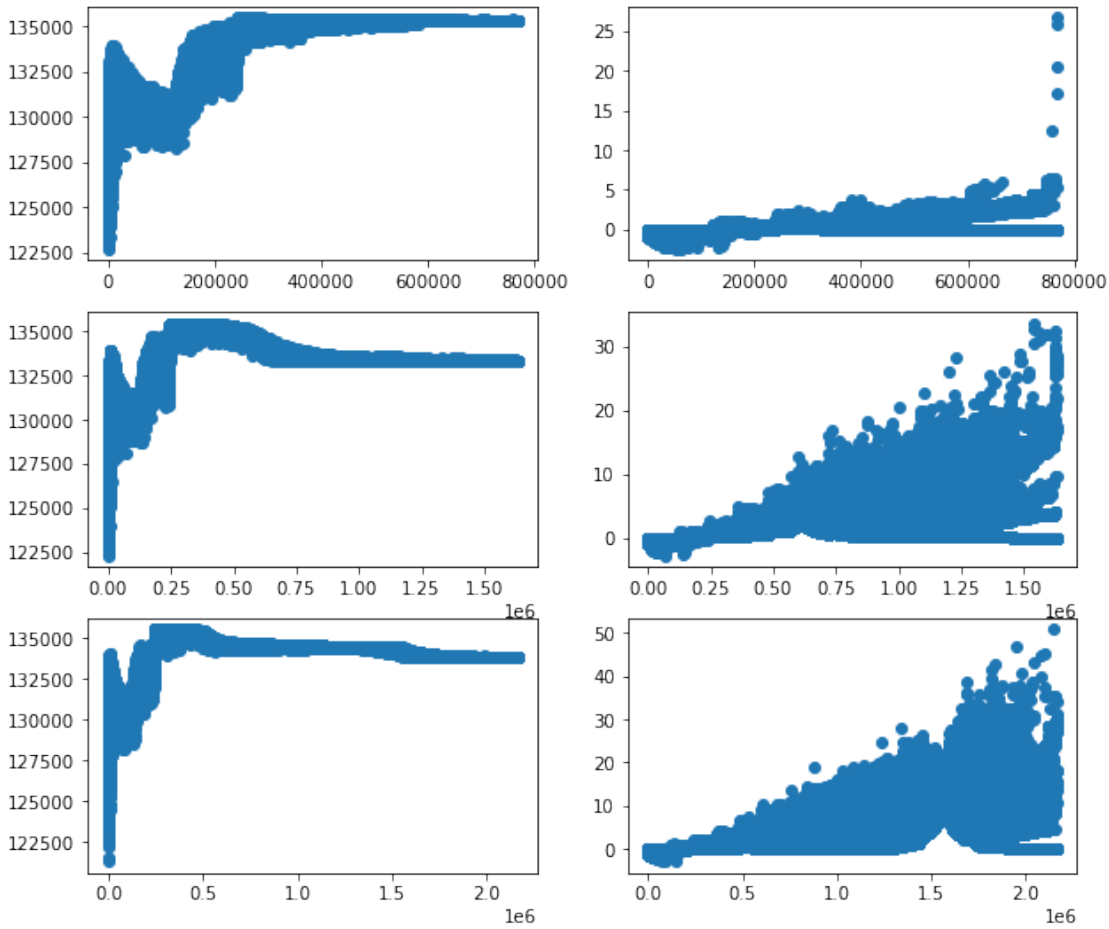
```
0.5353, 0.5353,
        0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353,
        0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5929])
finalReturns:  tensor([3.2408, 3.2984, 3.2672, 3.1741, 3.0373, 2.8692, 2.6784,
2.4711, 2.2517,
        2.0233, 1.7883, 1.5484, 1.3049, 1.0586, 0.8103, 0.5604, 0.3094])
----------------------------------------
iter  2  stage  7  ep  73532   adversary:  AdversaryModes.imitation_132
  actions:  tensor([24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24,
24, 24, 24,
        24, 24, 24, 24, 24, 24,  0])
loss=  tensor(0.0306, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.1997,
6.1997, 6.1997, 6.1997, 6.1997, 6.1997, 6.1997, 6.1997, 5.6068,
        5.1027, 4.6605, 4.2620, 3.8949, 3.5503, 3.2223, 2.9065, 2.5995, 2.2992,
        2.0038, 1.7120, 1.4230, 1.1360, 0.8506, 0.5663, 0.2828]) return=
133755.66666666666
probs of actions:  tensor([0.9977, 0.9982, 0.9975, 0.9975, 0.9984, 0.9978,
0.9975, 0.9990, 0.9988,
        0.9988, 0.9991, 0.9994, 0.9995, 0.9995, 0.9999, 0.9995, 0.9991, 0.9989,
        0.9993, 0.9994, 0.9993, 0.9992, 0.9997, 0.9996, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4536, 0.5567, 0.5300, 0.5366, 0.5350, 0.5354, 0.5353,
0.5353, 0.5353,
        0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353,
        0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5929])
finalReturns:  tensor([3.4933, 3.5509, 3.5197, 3.4266, 3.2898, 3.1216, 2.9309,
2.7236, 2.5041,
        2.2758, 2.0408, 1.8009, 1.5574, 1.3111, 1.0628, 0.8129, 0.5619, 0.3101])
----------------------------------------
iter  2  stage  6  ep  44880   adversary:  AdversaryModes.imitation_132
  actions:  tensor([24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24,
24, 24, 24,
        24, 24, 24, 24, 24, 24,  0])
loss=  tensor(0.0163, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.4820,
6.4820, 6.4820, 6.4820, 6.4820, 6.4820, 6.4820, 5.8892, 5.3851,
        4.9428, 4.5444, 4.1772, 3.8327, 3.5047, 3.1888, 2.8818, 2.5815, 2.2861,
        1.9944, 1.7053, 1.4184, 1.1329, 0.8486, 0.5652, 0.2823]) return=
133755.66666666666
probs of actions:  tensor([0.9986, 0.9989, 0.9985, 0.9985, 0.9991, 0.9987,
0.9990, 0.9997, 0.9994,
        0.9994, 0.9995, 0.9997, 1.0000, 0.9998, 1.0000, 0.9999, 0.9996, 0.9995,
        0.9997, 0.9998, 0.9997, 0.9997, 1.0000, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4536, 0.5567, 0.5300, 0.5366, 0.5350, 0.5354, 0.5353,
0.5353, 0.5353,
        0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353,
        0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5929])
finalReturns:  tensor([3.7462, 3.8038, 3.7726, 3.6796, 3.5428, 3.3746, 3.1838,
```

```
2.9765, 2.7571,
        2.5288, 2.2938, 2.0539, 1.8103, 1.5641, 1.3157, 1.0659, 0.8149, 0.5630,
        0.3106])
----------------------------------------
iter  2  stage  5  ep  5209   adversary:  AdversaryModes.imitation_132
  actions:  tensor([24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24,
24, 24, 24,
        24, 24, 24, 24, 24, 24,  0])
loss=  tensor(0.0176, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.7641,
6.7641, 6.7641, 6.7641, 6.7641, 6.7641, 6.1711, 5.6670, 5.2248,
        4.8263, 4.4592, 4.1147, 3.7867, 3.4708, 3.1638, 2.8635, 2.5681, 2.2763,
        1.9873, 1.7003, 1.4149, 1.1306, 0.8471, 0.5643, 0.2820]) return=
133755.66666666666
probs of actions:  tensor([0.9988, 0.9991, 0.9987, 0.9987, 0.9992, 0.9990,
0.9991, 0.9998, 0.9995,
        0.9995, 0.9996, 0.9998, 1.0000, 0.9999, 1.0000, 0.9999, 0.9997, 0.9995,
        0.9998, 0.9998, 0.9998, 0.9998, 1.0000, 0.9999, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4536, 0.5567, 0.5300, 0.5366, 0.5350, 0.5354, 0.5353,
0.5353, 0.5353,
        0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353,
        0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5929])
finalReturns:  tensor([3.9996, 4.0572, 4.0260, 3.9329, 3.7961, 3.6279, 3.4371,
3.2299, 3.0104,
        2.7821, 2.5471, 2.3072, 2.0637, 1.8174, 1.5691, 1.3192, 1.0682, 0.8164,
        0.5639, 0.3109])
----------------------------------------
iter  2  stage  4  ep  0   adversary:  AdversaryModes.imitation_132
  actions:  tensor([24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24,
24, 24, 24,
        24, 24, 24, 24, 24, 24,  0])
loss=  tensor(0.0217, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.0455,
7.0455, 7.0455, 7.0455, 7.0455, 6.4529, 5.9487, 5.5065, 5.1080,
        4.7409, 4.3964, 4.0684, 3.7525, 3.4455, 3.1452, 2.8498, 2.5580, 2.2690,
        1.9821, 1.6966, 1.4123, 1.1288, 0.8460, 0.5637, 0.2817]) return=
133755.66666666666
probs of actions:  tensor([0.9988, 0.9991, 0.9987, 0.9987, 0.9992, 0.9990,
0.9991, 0.9998, 0.9995,
        0.9995, 0.9996, 0.9998, 1.0000, 0.9999, 1.0000, 0.9999, 0.9997, 0.9995,
        0.9998, 0.9998, 0.9998, 0.9998, 1.0000, 0.9999, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4536, 0.5567, 0.5300, 0.5366, 0.5350, 0.5354, 0.5353,
0.5353, 0.5353,
        0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353,
        0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5929])
finalReturns:  tensor([4.2532, 4.3108, 4.2796, 4.1865, 4.0497, 3.8815, 3.6907,
3.4834, 3.2640,
        3.0357, 2.8007, 2.5608, 2.3173, 2.0710, 1.8226, 1.5728, 1.3218, 1.0700,
```

```
        0.8175, 0.5645, 0.3112])
---------------------------------------
iter  2  stage  3  ep  357   adversary:  AdversaryModes.imitation_132
  actions:  tensor([24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24,
24, 24, 24,
        24, 24, 24, 24, 24, 24,  0])
loss=  tensor(0.0256, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.3283,
7.3283, 7.3283, 7.3283, 6.7341, 6.2303, 5.7880, 5.3895, 5.0224,
        4.6779, 4.3499, 4.0340, 3.7270, 3.4267, 3.1313, 2.8396, 2.5505, 2.2636,
        1.9781, 1.6938, 1.4103, 1.1275, 0.8452, 0.5632, 0.2815]) return=
133755.66666666666
probs of actions:  tensor([0.9988, 0.9991, 0.9987, 0.9990, 0.9994, 0.9990,
0.9991, 0.9998, 0.9995,
        0.9995, 0.9996, 0.9998, 1.0000, 0.9999, 1.0000, 0.9999, 0.9997, 0.9996,
        0.9998, 0.9999, 0.9998, 0.9998, 1.0000, 0.9999, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4536, 0.5567, 0.5300, 0.5366, 0.5350, 0.5354, 0.5353,
0.5353, 0.5353,
        0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353,
        0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5929])
finalReturns:  tensor([4.5069, 4.5645, 4.5334, 4.4403, 4.3035, 4.1353, 3.9445,
3.7372, 3.5178,
        3.2895, 3.0545, 2.8146, 2.5711, 2.3248, 2.0764, 1.8266, 1.5756, 1.3238,
        1.0713, 0.8183, 0.5650, 0.3114])
---------------------------------------
iter  2  stage  2  ep  216   adversary:  AdversaryModes.imitation_132
  actions:  tensor([24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24,
24, 24, 24,
        24, 24, 24, 24, 24, 24,  0])
loss=  tensor(0.0289, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.6043,
7.6043, 7.6043, 7.0167, 6.5114, 6.0694, 5.6709, 5.3037, 4.9592,
        4.6312, 4.3153, 4.0084, 3.7081, 3.4127, 3.1209, 2.8319, 2.5449, 2.2595,
        1.9752, 1.6917, 1.4089, 1.1265, 0.8446, 0.5629, 0.2814]) return=
133755.66666666666
probs of actions:  tensor([0.9989, 0.9991, 0.9990, 0.9991, 0.9995, 0.9991,
0.9992, 0.9998, 0.9995,
        0.9996, 0.9997, 0.9999, 1.0000, 0.9999, 1.0000, 1.0000, 0.9997, 0.9996,
        0.9998, 0.9999, 0.9998, 0.9998, 1.0000, 0.9999, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4536, 0.5567, 0.5300, 0.5366, 0.5350, 0.5354, 0.5353,
0.5353, 0.5353,
        0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353,
        0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5929])
finalReturns:  tensor([4.7610, 4.8186, 4.7873, 4.6942, 4.5574, 4.3893, 4.1985,
3.9912, 3.7718,
        3.5434, 3.3084, 3.0685, 2.8250, 2.5787, 2.3304, 2.0805, 1.8295, 1.5777,
        1.3252, 1.0723, 0.8189, 0.5653, 0.3115])
---------------------------------------
```

```
iter  2  stage  1  ep  0   adversary:  AdversaryModes.imitation_132
  actions:  tensor([24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24,
24, 24, 24,
        24, 24, 24, 24, 24, 24,  0])
loss=  tensor(0.0342, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.9073,
7.9073, 7.2930, 6.7938, 6.3504, 5.9522, 5.5850, 5.2405, 4.9125,
        4.5966, 4.2896, 3.9893, 3.6939, 3.4021, 3.1131, 2.8262, 2.5407, 2.2564,
        1.9729, 1.6901, 1.4078, 1.1258, 0.8441, 0.5626, 0.2812]) return=
133755.66666666666
probs of actions:  tensor([0.9989, 0.9991, 0.9990, 0.9991, 0.9995, 0.9991,
0.9992, 0.9998, 0.9995,
        0.9996, 0.9997, 0.9999, 1.0000, 0.9999, 1.0000, 1.0000, 0.9997, 0.9996,
        0.9998, 0.9999, 0.9998, 0.9998, 1.0000, 0.9999, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4536, 0.5567, 0.5300, 0.5366, 0.5350, 0.5354, 0.5353,
0.5353, 0.5353,
        0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353,
        0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5929])
finalReturns:  tensor([5.0147, 5.0723, 5.0415, 4.9482, 4.8115, 4.6433, 4.4525,
4.2452, 4.0258,
        3.7975, 3.5625, 3.3226, 3.0791, 2.8328, 2.5844, 2.3346, 2.0836, 1.8318,
        1.5793, 1.3263, 1.0730, 0.8194, 0.5656, 0.3117])
----------------------------------------
iter  2  stage  0  ep  64   adversary:  AdversaryModes.imitation_132
  actions:  tensor([24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24,
24, 24, 24,
        24, 24, 24, 24, 24, 24,  0])
loss=  tensor(0.0387, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.1053,
7.5941, 7.0703, 6.6326, 6.2331, 5.8662, 5.5216, 5.1936, 4.8777,
        4.5708, 4.2705, 3.9750, 3.6833, 3.3943, 3.1073, 2.8219, 2.5375, 2.2541,
        1.9713, 1.6889, 1.4070, 1.1253, 0.8437, 0.5624, 0.2812]) return=
133755.66666666666
probs of actions:  tensor([0.9990, 0.9992, 0.9991, 0.9992, 0.9995, 0.9991,
0.9992, 0.9998, 0.9996,
        0.9996, 0.9997, 0.9999, 1.0000, 0.9999, 1.0000, 1.0000, 0.9997, 0.9996,
        0.9998, 0.9999, 0.9998, 0.9998, 1.0000, 0.9999, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4536, 0.5567, 0.5300, 0.5366, 0.5350, 0.5354, 0.5353,
0.5353, 0.5353,
        0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353,
        0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5353, 0.5929])
finalReturns:  tensor([5.2702, 5.3278, 5.2950, 5.2026, 5.0656, 4.8975, 4.7067,
4.4994, 4.2800,
        4.0516, 3.8166, 3.5768, 3.3332, 3.0869, 2.8386, 2.5887, 2.3378, 2.0859,
        1.8334, 1.5805, 1.3271, 1.0735, 0.8198, 0.5658, 0.3117])
0,[1e-05,1][1, 10000, 1, 1],1682794213 saved
[2164266, 'tensor([0., 0., 0., 1., 0., 0., 0., 0., 0., 0., 0., 0.])',
133755.66666666666, 78214.06666666668, 0.038705166429281235, 1e-05, 1, 0,
```

'tensor([24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24,
24,\n   24, 24, 24, 24, 24, 24, 0])', '[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.\n 1.]', '0,[1e-05,1][1, 10000, 1,
1],1682794213', 25, 50, 170327.33349408704, 211165.96084616287,
87190.73961268678, 133755.6666666667, 130787.53866666667, 82417.65584066519,
82417.65584066519, 99036.07704587854, 99036.07704587854, 100843.49835063974,
82417.65584066519, 99036.07704587854]



policy reset
----
iter 0 stage 24 ep 99999 adversary: AdversaryModes.imitation_128
  actions: tensor([0, 0, 0, 0, 0, 7, 0, 0, 0, 5, 0, 4, 1, 0, 0, 1, 0, 0, 2, 0,
0, 0, 0, 1,
        0])
loss= tensor(-0., grad_fn=<NegBackward0>) , base rewards= tensor([0.5134,
0.5134, 0.5134, 0.5134, 0.5134, 0.5134, 0.5134, 0.5134, 0.5134,
        0.5134, 0.5134, 0.5134, 0.5134, 0.5134, 0.5134, 0.5134, 0.5134, 0.5134,
        0.5134, 0.5134, 0.5134, 0.5134, 0.5134, 0.5134, 0.5134]) return=
127976.82117179644

```
probs of actions:  tensor([0.8485, 0.8295, 0.8628, 0.8631, 0.7988, 0.0024,
0.8453, 0.8518, 0.8468,
        0.0049, 0.8436, 0.0070, 0.0926, 0.8624, 0.8310, 0.0958, 0.8208, 0.8386,
        0.0177, 0.8306, 0.8285, 0.8579, 0.8012, 0.0884, 0.9670],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5112, 0.5094, 0.5099, 0.5098, 0.5098, 0.5049, 0.5351,
0.5036, 0.5114,
        0.5069, 0.5279, 0.5037, 0.5252, 0.5095, 0.5099, 0.5097, 0.5134, 0.5089,
        0.5096, 0.5169, 0.5080, 0.5102, 0.5097, 0.5097, 0.5134])
finalReturns:  tensor([0.])
----------------------------------------
iter  0  stage  23  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([16,  0, 16, 15, 14, 15, 11,  3,  0,  0, 20, 15, 15,  1, 11,
8,  8, 16,
         1,  1, 15, 16,  0, 15,  0])
loss=  tensor(0.0781, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.0114,
1.0114, 1.0114, 1.0114, 1.0114, 1.0114, 1.0114, 1.0114, 1.0114,
        1.0114, 1.0114, 1.0114, 1.0114, 1.0114, 1.0114, 1.0114, 1.0114, 1.0114,
        1.0114, 1.0114, 1.0114, 1.0114, 1.0114, 1.0114, 0.5126]) return=
131016.7267856707
probs of actions:  tensor([0.0725, 0.2308, 0.0930, 0.1020, 0.0359, 0.0781,
0.0610, 0.0218, 0.3880,
        0.3613, 0.0086, 0.1072, 0.0867, 0.0801, 0.0709, 0.0041, 0.0056, 0.0781,
        0.0742, 0.0698, 0.0922, 0.0806, 0.1969, 0.0934, 0.9848],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4856, 0.5681, 0.4701, 0.5498, 0.5293, 0.5285, 0.5421,
0.5377, 0.5134,
        0.5089, 0.4700, 0.5611, 0.5236, 0.5553, 0.4901, 0.5454, 0.5218, 0.5085,
        0.5621, 0.5005, 0.4932, 0.5376, 0.5548, 0.4763, 0.5677])
finalReturns:  tensor([0.0326, 0.0551])
----------------------------------------
iter  0  stage  22  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([22, 22, 18, 15, 20, 18, 12, 12, 27, 19, 15, 18, 15, 12,  1,
22,  0, 18,
        15, 14, 18, 18, 21, 22,  0])
loss=  tensor(0.8808, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.4995,
1.4995, 1.4995, 1.4995, 1.4995, 1.4995, 1.4995, 1.4995, 1.4995,
        1.4995, 1.4995, 1.4995, 1.4995, 1.4995, 1.4995, 1.4995, 1.4995, 1.4995,
        1.4995, 1.4995, 1.4995, 1.4995, 1.4995, 0.9377, 0.4405]) return=
131827.89335068475
probs of actions:  tensor([0.0964, 0.0982, 0.3525, 0.0528, 0.0570, 0.3105,
0.1858, 0.0708, 0.0098,
        0.0618, 0.0573, 0.3612, 0.0493, 0.1349, 0.0565, 0.1031, 0.0742, 0.2706,
        0.0473, 0.0258, 0.3585, 0.2952, 0.0171, 0.0871, 0.9850],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5426, 0.5381, 0.5380, 0.5118, 0.5403, 0.5456,
0.5264, 0.4727,
        0.5651, 0.5343, 0.5203, 0.5425, 0.5363, 0.5430, 0.4568, 0.5926, 0.4577,
```

```
        0.5589, 0.5271, 0.5191, 0.5329, 0.5177, 0.5256, 0.5747])
finalReturns:  tensor([0.1185, 0.1626, 0.1342])
----------------------------------------
iter  0  stage  21  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([22, 22, 22,  0, 22, 22, 18, 22, 18, 22, 23, 22, 17, 22, 14,
22, 23, 22,
        22, 16, 22, 22, 16, 22,  0])
loss=  tensor(1.5077, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.9284,
1.9284, 1.9284, 1.9284, 1.9284, 1.9284, 1.9284, 1.9284, 1.9284,
        1.9284, 1.9284, 1.9284, 1.9284, 1.9284, 1.9284, 1.9284, 1.9284, 1.9284,
        1.9284, 1.9284, 1.9284, 1.9284, 1.3481, 0.8552, 0.4170]) return=
131625.96944387257
probs of actions:  tensor([0.5168, 0.4888, 0.5555, 0.0028, 0.5101, 0.5398,
0.1094, 0.5222, 0.1204,
        0.5231, 0.1283, 0.5068, 0.0124, 0.5243, 0.0104, 0.5004, 0.1249, 0.4907,
        0.5293, 0.0305, 0.5330, 0.6392, 0.0121, 0.5230, 0.9884],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4628, 0.5426, 0.5221, 0.5756, 0.4456, 0.5473, 0.5369,
0.5124, 0.5456,
        0.5102, 0.5257, 0.5290, 0.5450, 0.5075, 0.5597, 0.4951, 0.5296, 0.5280,
        0.5257, 0.5491, 0.5036, 0.5319, 0.5475, 0.5040, 0.5802])
finalReturns:  tensor([0.2352, 0.2836, 0.2289, 0.1632])
----------------------------------------
iter  0  stage  20  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([23, 22, 22, 22, 22, 22, 22, 23, 19, 22, 27, 22, 22, 22, 23,
23, 23, 18,
        22, 22, 22, 23, 20, 22,  0])
loss=  tensor(2.8372, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.2744,
2.2744, 2.2744, 2.2744, 2.2744, 2.2744, 2.2744, 2.2744, 2.2744,
        2.2744, 2.2744, 2.2744, 2.2744, 2.2744, 2.2744, 2.2744, 2.2744, 2.2744,
        2.2744, 2.2744, 2.2744, 1.7008, 1.2064, 0.7685, 0.3738]) return=
131444.76673770836
probs of actions:  tensor([0.1846, 0.6751, 0.7233, 0.6721, 0.6664, 0.6970,
0.6251, 0.1851, 0.0142,
        0.6894, 0.0127, 0.6822, 0.6845, 0.6798, 0.1922, 0.1971, 0.1872, 0.0422,
        0.6683, 0.6213, 0.7433, 0.1346, 0.0178, 0.7598, 0.9964],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4583, 0.5464, 0.5212, 0.5274, 0.5259, 0.5262, 0.5261,
0.5217, 0.5423,
        0.5139, 0.5047, 0.5445, 0.5216, 0.5273, 0.5214, 0.5255, 0.5245, 0.5453,
        0.5103, 0.5302, 0.5252, 0.5219, 0.5383, 0.5177, 0.5767])
finalReturns:  tensor([0.4053, 0.4537, 0.4263, 0.3258, 0.2029])
----------------------------------------
iter  0  stage  19  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([23, 22, 23, 22, 23, 22, 23, 22, 22, 22, 23, 18, 23, 23, 22,
22, 23, 23,
        22, 22, 22, 22, 22, 19,  0])
loss=  tensor(2.8711, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.6216,
```

```
2.6216, 2.6216, 2.6216, 2.6216, 2.6216, 2.6216, 2.6216, 2.6216,
        2.6216, 2.6216, 2.6216, 2.6216, 2.6216, 2.6216, 2.6216, 2.6216, 2.6216,
        2.6216, 2.6216, 2.0477, 1.5533, 1.1155, 0.7175, 0.3483]) return=
131442.16007456576
probs of actions:  tensor([0.2345, 0.7010, 0.2029, 0.6672, 0.2341, 0.7003,
0.2802, 0.7011, 0.6640,
        0.7034, 0.2696, 0.0222, 0.2170, 0.2349, 0.6298, 0.6698, 0.2356, 0.2339,
        0.6703, 0.6671, 0.7788, 0.8038, 0.7787, 0.0031, 0.9995],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4583, 0.5464, 0.5167, 0.5312, 0.5204, 0.5303, 0.5206,
0.5302, 0.5252,
        0.5264, 0.5216, 0.5460, 0.5057, 0.5295, 0.5280, 0.5257, 0.5218, 0.5254,
        0.5290, 0.5255, 0.5263, 0.5261, 0.5262, 0.5385, 0.5633])
finalReturns:  tensor([0.5842, 0.6326, 0.6007, 0.5124, 0.3842, 0.2149])
----------------------------------------
iter  0  stage  18  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([23, 23, 22, 23, 22, 22, 23, 22, 22, 23, 22, 22, 22, 22, 22,
23, 22, 22,
        27, 23, 23, 22, 22, 22,  0])
loss=  tensor(5.6294, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.9041,
2.9041, 2.9041, 2.9041, 2.9041, 2.9041, 2.9041, 2.9041, 2.9041,
        2.9041, 2.9041, 2.9041, 2.9041, 2.9041, 2.9041, 2.9041, 2.9041, 2.9041,
        2.9041, 2.3293, 1.8351, 1.4136, 1.0266, 0.6685, 0.3275]) return=
131314.67447750745
probs of actions:  tensor([0.3571, 0.3441, 0.6513, 0.3953, 0.5974, 0.6013,
0.4182, 0.6063, 0.5660,
        0.3563, 0.5332, 0.6063, 0.6232, 0.6000, 0.5383, 0.3729, 0.6065, 0.6052,
        0.0218, 0.4118, 0.2705, 0.7124, 0.6872, 0.7186, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4583, 0.5419, 0.5249, 0.5220, 0.5299, 0.5252, 0.5219,
0.5299, 0.5252,
        0.5219, 0.5299, 0.5252, 0.5264, 0.5261, 0.5262, 0.5217, 0.5300, 0.5252,
        0.5019, 0.5407, 0.5207, 0.5302, 0.5252, 0.5264, 0.5745])
finalReturns:  tensor([0.8156, 0.8885, 0.8419, 0.7427, 0.5995, 0.4324, 0.2470])
----------------------------------------
iter  0  stage  17  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([23, 23, 23, 23, 23, 23, 23, 22, 23, 23, 22, 23, 22, 20, 23,
22, 23, 23,
        23, 23, 23, 23, 22, 23,  0])
loss=  tensor(2.5241, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.2482,
3.2482, 3.2482, 3.2482, 3.2482, 3.2482, 3.2482, 3.2482, 3.2482,
        3.2482, 3.2482, 3.2482, 3.2482, 3.2482, 3.2482, 3.2482, 3.2482, 3.2482,
        2.6695, 2.1762, 1.7414, 1.3482, 0.9849, 0.6432, 0.3172]) return=
131291.9557196191
probs of actions:  tensor([0.7267, 0.7023, 0.7028, 0.7538, 0.6919, 0.7213,
0.7498, 0.2223, 0.7321,
        0.7212, 0.1687, 0.7095, 0.2447, 0.0093, 0.6804, 0.2080, 0.7106, 0.7318,
        0.7188, 0.7547, 0.6970, 0.6964, 0.2844, 0.7467, 1.0000],
```

```
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4583, 0.5419, 0.5204, 0.5258, 0.5244, 0.5248, 0.5247,
0.5292, 0.5209,
        0.5257, 0.5290, 0.5210, 0.5301, 0.5336, 0.5144, 0.5318, 0.5203, 0.5258,
        0.5244, 0.5248, 0.5247, 0.5247, 0.5292, 0.5209, 0.5786])
finalReturns:  tensor([1.0049, 1.0578, 1.0266, 0.9367, 0.8051, 0.6438, 0.4563,
0.2614])
----------------------------------------
iter  0  stage  16  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([23, 23, 23, 23, 23, 23, 27, 23, 23, 23, 23, 27, 22, 23, 23,
23, 22, 23,
        23, 23, 23, 23, 23, 23,  0])
loss=  tensor(4.1050, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.5591,
3.5591, 3.5591, 3.5591, 3.5591, 3.5591, 3.5591, 3.5591, 3.5591,
        3.5591, 3.5591, 3.5591, 3.5591, 3.5591, 3.5591, 3.5591, 3.5591, 2.9813,
        2.4879, 2.0498, 1.6550, 1.2904, 0.9477, 0.6211, 0.3062]) return=
131053.19988043512
probs of actions:  tensor([0.7668, 0.7477, 0.7517, 0.7799, 0.7302, 0.7619,
0.0950, 0.7659, 0.7621,
        0.7609, 0.7681, 0.0614, 0.1864, 0.7542, 0.6877, 0.7610, 0.1592, 0.7650,
        0.7330, 0.7773, 0.7630, 0.7196, 0.7487, 0.8173, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4583, 0.5419, 0.5204, 0.5258, 0.5244, 0.5248, 0.5047,
0.5400, 0.5209,
        0.5257, 0.5245, 0.5048, 0.5445, 0.5171, 0.5266, 0.5242, 0.5293, 0.5209,
        0.5257, 0.5245, 0.5248, 0.5247, 0.5247, 0.5247, 0.5776])
finalReturns:  tensor([1.2177, 1.2661, 1.2387, 1.1511, 1.0214, 0.8613, 0.6793,
0.4812, 0.2714])
----------------------------------------
iter  0  stage  15  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([23, 23, 23, 27, 27, 27, 23, 23, 27, 23, 27, 23, 23, 27, 27,
23, 27, 27,
        27, 27, 27, 23, 23, 22,  0])
loss=  tensor(9.6563, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.7219,
3.7219, 3.7219, 3.7219, 3.7219, 3.7219, 3.7219, 3.7219, 3.7219,
        3.7219, 3.7219, 3.7219, 3.7219, 3.7219, 3.7219, 3.7219, 3.1328, 2.6420,
        2.2066, 1.8260, 1.4805, 1.1610, 0.8602, 0.5731, 0.2854]) return=
130253.95806091337
probs of actions:  tensor([0.5601, 0.5729, 0.5629, 0.4018, 0.3564, 0.3282,
0.4954, 0.5843, 0.4051,
        0.5535, 0.4352, 0.5639, 0.5251, 0.3924, 0.5742, 0.5398, 0.3543, 0.4443,
        0.5178, 0.4116, 0.2818, 0.5405, 0.6194, 0.1406, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4583, 0.5419, 0.5204, 0.5058, 0.5197, 0.5162, 0.5371,
0.5216, 0.5055,
        0.5398, 0.5010, 0.5410, 0.5207, 0.5057, 0.5197, 0.5362, 0.5018, 0.5207,
        0.5160, 0.5172, 0.5169, 0.5369, 0.5217, 0.5300, 0.5736])
finalReturns:  tensor([1.5490, 1.6019, 1.5909, 1.5056, 1.3702, 1.1986, 1.0012,
```

```
0.7650, 0.5304,
        0.2882])
---------------------------------------
iter  0  stage  14  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([27, 27, 23, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,
27, 27, 27,
        27, 27, 27, 22, 27, 23,  0])
loss=  tensor(5.8675, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.9442,
3.9442, 3.9442, 3.9442, 3.9442, 3.9442, 3.9442, 3.9442, 3.9442,
        3.9442, 3.9442, 3.9442, 3.9442, 3.9442, 3.9442, 3.3544, 2.8637, 2.4413,
        2.0669, 1.7265, 1.4106, 1.1124, 0.8272, 0.5516, 0.2698]) return=
129554.93117643229
probs of actions:  tensor([0.8762, 0.8405, 0.1034, 0.9194, 0.8752, 0.8740,
0.9195, 0.8534, 0.9118,
        0.8799, 0.9213, 0.8607, 0.8698, 0.9070, 0.9539, 0.9097, 0.9026, 0.9266,
        0.9496, 0.9206, 0.8964, 0.0236, 0.8098, 0.2168, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5375, 0.5318, 0.5029, 0.5205, 0.5160, 0.5171,
0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5414, 0.4979, 0.5417, 0.5734])
finalReturns:  tensor([1.8287, 1.9016, 1.8754, 1.7808, 1.6383, 1.4618, 1.2608,
1.0420, 0.7858,
        0.5636, 0.3035])
---------------------------------------
iter  0  stage  13  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,
27, 27, 27,
        27, 27, 27, 27, 27, 27,  0])
loss=  tensor(0.3084, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.1841,
4.1841, 4.1841, 4.1841, 4.1841, 4.1841, 4.1841, 4.1841, 4.1841,
        4.1841, 4.1841, 4.1841, 4.1841, 4.1841, 3.5943, 3.1036, 2.6813, 2.3068,
        1.9664, 1.6505, 1.3523, 1.0671, 0.7915, 0.5229, 0.2595]) return=
129338.69866666665
probs of actions:  tensor([0.9714, 0.9610, 0.9711, 0.9835, 0.9692, 0.9713,
0.9829, 0.9650, 0.9811,
        0.9735, 0.9837, 0.9670, 0.9692, 0.9851, 0.9930, 0.9845, 0.9812, 0.9859,
        0.9918, 0.9852, 0.9800, 0.9791, 0.9566, 0.9344, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5375, 0.5118, 0.5182, 0.5166, 0.5170, 0.5169,
0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5898])
finalReturns:  tensor([2.0919, 2.1648, 2.1385, 2.0440, 1.9015, 1.7249, 1.5239,
1.3052, 1.0735,
        0.8322, 0.5839, 0.3303])
---------------------------------------
iter  0  stage  12  ep  99999   adversary:  AdversaryModes.imitation_128
```

```
  actions:  tensor([27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,
27, 27, 27,
        27, 27, 27, 27, 27, 27,  0])
loss=  tensor(0.0437, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.4407,
4.4407, 4.4407, 4.4407, 4.4407, 4.4407, 4.4407, 4.4407, 4.4407,
        4.4407, 4.4407, 4.4407, 4.4407, 3.8509, 3.3602, 2.9378, 2.5634, 2.2230,
        1.9071, 1.6089, 1.3237, 1.0481, 0.7795, 0.5161, 0.2566]) return=
129338.69866666665
probs of actions:  tensor([0.9954, 0.9937, 0.9958, 0.9977, 0.9947, 0.9956,
0.9974, 0.9943, 0.9972,
        0.9960, 0.9976, 0.9947, 0.9961, 0.9984, 0.9995, 0.9984, 0.9975, 0.9986,
        0.9997, 0.9986, 0.9983, 0.9978, 0.9940, 0.9915, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5375, 0.5118, 0.5182, 0.5166, 0.5170, 0.5169,
0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5898])
finalReturns:  tensor([2.3522, 2.4251, 2.3989, 2.3043, 2.1618, 1.9853, 1.7843,
1.5655, 1.3338,
        1.0925, 0.8442, 0.5907, 0.3332])
----------------------------------------
iter  0  stage  11  ep  67016   adversary:  AdversaryModes.imitation_128
  actions:  tensor([27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,
27, 27, 27,
        27, 27, 27, 27, 27, 27,  0])
loss=  tensor(0.0133, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.6951,
4.6951, 4.6951, 4.6951, 4.6951, 4.6951, 4.6951, 4.6951, 4.6951,
        4.6951, 4.6951, 4.6951, 4.1053, 3.6146, 3.1923, 2.8179, 2.4775, 2.1615,
        1.8634, 1.5782, 1.3025, 1.0339, 0.7705, 0.5110, 0.2544]) return=
129338.69866666665
probs of actions:  tensor([0.9983, 0.9978, 0.9985, 0.9992, 0.9981, 0.9985,
0.9991, 0.9980, 0.9990,
        0.9986, 0.9991, 0.9990, 0.9989, 0.9997, 1.0000, 0.9998, 0.9992, 0.9999,
        1.0000, 1.0000, 0.9999, 0.9993, 0.9979, 0.9973, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5375, 0.5118, 0.5182, 0.5166, 0.5170, 0.5169,
0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5898])
finalReturns:  tensor([2.6147, 2.6876, 2.6614, 2.5668, 2.4243, 2.2478, 2.0468,
1.8280, 1.5963,
        1.3550, 1.1067, 0.8532, 0.5957, 0.3354])
----------------------------------------
iter  0  stage  10  ep  0   adversary:  AdversaryModes.imitation_128
  actions:  tensor([27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,
27, 27, 27,
        27, 27, 27, 27, 27, 27,  0])
loss=  tensor(0.0165, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.9480,
```

```
4.9480, 4.9480, 4.9480, 4.9480, 4.9480, 4.9480, 4.9480, 4.9480,
        4.9480, 4.9480, 4.3581, 3.8674, 3.4451, 3.0707, 2.7303, 2.4144, 2.1162,
        1.8310, 1.5553, 1.2867, 1.0233, 0.7638, 0.5073, 0.2528]) return=
129338.69866666665
probs of actions:  tensor([0.9983, 0.9978, 0.9985, 0.9992, 0.9981, 0.9985,
0.9991, 0.9980, 0.9990,
        0.9986, 0.9991, 0.9990, 0.9989, 0.9997, 1.0000, 0.9998, 0.9992, 0.9999,
        1.0000, 1.0000, 0.9999, 0.9993, 0.9979, 0.9973, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5375, 0.5118, 0.5182, 0.5166, 0.5170, 0.5169,
0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5898])
finalReturns:  tensor([2.8788, 2.9517, 2.9255, 2.8309, 2.6884, 2.5119, 2.3109,
2.0921, 1.8604,
        1.6191, 1.3708, 1.1173, 0.8598, 0.5995, 0.3370])
----------------------------------------
iter  0  stage  9  ep  11308   adversary:  AdversaryModes.imitation_128
  actions:  tensor([27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,
27, 27, 27,
        27, 27, 27, 27, 27, 27,  0])
loss=  tensor(0.0184, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.1996,
5.1996, 5.1996, 5.1996, 5.1996, 5.1996, 5.1996, 5.1996, 5.1996,
        5.1996, 4.6098, 4.1191, 3.6967, 3.3223, 2.9819, 2.6660, 2.3678, 2.0826,
        1.8069, 1.5383, 1.2750, 1.0155, 0.7589, 0.5044, 0.2516]) return=
129338.69866666665
probs of actions:  tensor([0.9984, 0.9979, 0.9986, 0.9992, 0.9982, 0.9985,
0.9991, 0.9981, 0.9990,
        0.9990, 0.9993, 0.9991, 0.9990, 0.9997, 1.0000, 0.9998, 0.9992, 1.0000,
        1.0000, 1.0000, 0.9999, 0.9994, 0.9980, 0.9975, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5375, 0.5118, 0.5182, 0.5166, 0.5170, 0.5169,
0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5898])
finalReturns:  tensor([3.1441, 3.2170, 3.1908, 3.0962, 2.9537, 2.7772, 2.5762,
2.3574, 2.1257,
        1.8844, 1.6361, 1.3826, 1.1251, 0.8648, 0.6023, 0.3382])
----------------------------------------
iter  0  stage  8  ep  0   adversary:  AdversaryModes.imitation_128
  actions:  tensor([27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,
27, 27, 27,
        27, 27, 27, 27, 27, 27,  0])
loss=  tensor(0.0226, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.4503,
5.4503, 5.4503, 5.4503, 5.4503, 5.4503, 5.4503, 5.4503, 5.4503,
        4.8605, 4.3698, 3.9474, 3.5730, 3.2326, 2.9167, 2.6185, 2.3333, 2.0576,
        1.7890, 1.5257, 1.2662, 1.0096, 0.7551, 0.5023, 0.2507]) return=
129338.69866666665
```

```
probs of actions:  tensor([0.9984, 0.9979, 0.9986, 0.9992, 0.9982, 0.9985,
0.9991, 0.9981, 0.9990,
        0.9990, 0.9993, 0.9991, 0.9990, 0.9997, 1.0000, 0.9998, 0.9992, 1.0000,
        1.0000, 1.0000, 0.9999, 0.9994, 0.9980, 0.9975, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5375, 0.5118, 0.5182, 0.5166, 0.5170, 0.5169,
0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5898])
finalReturns:  tensor([3.4103, 3.4832, 3.4570, 3.3624, 3.2199, 3.0434, 2.8424,
2.6236, 2.3919,
        2.1506, 1.9023, 1.6488, 1.3914, 1.1310, 0.8685, 0.6044, 0.3391])
----------------------------------------
iter  0  stage  7  ep  23219   adversary:  AdversaryModes.imitation_128
  actions:  tensor([27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,
27, 27, 27,
        27, 27, 27, 27, 27, 27,  0])
loss=  tensor(0.0149, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.7003,
5.7003, 5.7003, 5.7003, 5.7003, 5.7003, 5.7003, 5.7003, 5.1105,
        4.6198, 4.1974, 3.8230, 3.4826, 3.1667, 2.8685, 2.5833, 2.3077, 2.0391,
        1.7757, 1.5162, 1.2596, 1.0052, 0.7523, 0.5007, 0.2500]) return=
129338.69866666665
probs of actions:  tensor([0.9988, 0.9983, 0.9990, 0.9994, 0.9986, 0.9989,
0.9993, 0.9990, 1.0000,
        0.9993, 0.9995, 0.9994, 0.9998, 0.9998, 1.0000, 1.0000, 0.9996, 1.0000,
        1.0000, 1.0000, 1.0000, 0.9996, 0.9986, 0.9984, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5375, 0.5118, 0.5182, 0.5166, 0.5170, 0.5169,
0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5898])
finalReturns:  tensor([3.6772, 3.7501, 3.7239, 3.6293, 3.4868, 3.3103, 3.1093,
2.8905, 2.6588,
        2.4175, 2.1692, 1.9157, 1.6582, 1.3979, 1.1354, 0.8713, 0.6060, 0.3398])
----------------------------------------
iter  0  stage  6  ep  0   adversary:  AdversaryModes.imitation_128
  actions:  tensor([27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,
27, 27, 27,
        27, 27, 27, 27, 27, 27,  0])
loss=  tensor(0.0181, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.9498,
5.9498, 5.9498, 5.9498, 5.9498, 5.9498, 5.9498, 5.3600, 4.8693,
        4.4470, 4.0725, 3.7322, 3.4162, 3.1181, 2.8329, 2.5572, 2.2886, 2.0252,
        1.7657, 1.5091, 1.2547, 1.0019, 0.7503, 0.4996, 0.2495]) return=
129338.69866666665
probs of actions:  tensor([0.9988, 0.9983, 0.9990, 0.9994, 0.9986, 0.9989,
0.9993, 0.9990, 1.0000,
        0.9993, 0.9995, 0.9994, 0.9998, 0.9998, 1.0000, 1.0000, 0.9996, 1.0000,
        1.0000, 1.0000, 1.0000, 0.9996, 0.9986, 0.9984, 1.0000],
```

```
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5375, 0.5118, 0.5182, 0.5166, 0.5170, 0.5169,
0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5898])
finalReturns:  tensor([3.9446, 4.0175, 3.9913, 3.8967, 3.7542, 3.5777, 3.3767,
3.1579, 2.9262,
        2.6849, 2.4366, 2.1831, 1.9257, 1.6653, 1.4028, 1.1387, 0.8734, 0.6072,
        0.3403])
----------------------------------------
iter  0  stage  5  ep  63   adversary:  AdversaryModes.imitation_128
  actions:  tensor([27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,
27, 27, 27,
        27, 27, 27, 27, 27, 27,  0])
loss=  tensor(0.0223, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.1990,
6.1990, 6.1990, 6.1990, 6.1990, 6.1990, 5.6091, 5.1185, 4.6961,
        4.3217, 3.9813, 3.6654, 3.3672, 3.0820, 2.8063, 2.5377, 2.2744, 2.0149,
        1.7583, 1.5038, 1.2510, 0.9994, 0.7487, 0.4987, 0.2491]) return=
129338.69866666665
probs of actions:  tensor([0.9988, 0.9984, 0.9990, 0.9994, 0.9986, 0.9990,
0.9994, 0.9990, 1.0000,
        0.9994, 0.9995, 0.9995, 0.9998, 0.9998, 1.0000, 1.0000, 0.9996, 1.0000,
        1.0000, 1.0000, 1.0000, 0.9996, 0.9986, 0.9985, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5375, 0.5118, 0.5182, 0.5166, 0.5170, 0.5169,
0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5898])
finalReturns:  tensor([4.2124, 4.2853, 4.2591, 4.1645, 4.0220, 3.8455, 3.6445,
3.4257, 3.1940,
        2.9527, 2.7044, 2.4509, 2.1934, 1.9331, 1.6706, 1.4065, 1.1412, 0.8750,
        0.6081, 0.3407])
----------------------------------------
iter  0  stage  4  ep  250   adversary:  AdversaryModes.imitation_128
  actions:  tensor([27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,
27, 27, 27,
        27, 27, 27, 27, 27, 27,  0])
loss=  tensor(0.0241, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.4476,
6.4476, 6.4476, 6.4476, 6.4476, 5.8581, 5.3673, 4.9450, 4.5705,
        4.2302, 3.9142, 3.6161, 3.3309, 3.0552, 2.7866, 2.5232, 2.2637, 2.0071,
        1.7527, 1.4999, 1.2483, 0.9976, 0.7475, 0.4980, 0.2489]) return=
129338.69866666665
probs of actions:  tensor([0.9989, 0.9985, 0.9991, 0.9995, 0.9990, 0.9993,
0.9996, 0.9991, 1.0000,
        0.9994, 0.9995, 0.9995, 0.9998, 0.9998, 1.0000, 1.0000, 0.9997, 1.0000,
        1.0000, 1.0000, 1.0000, 0.9997, 0.9987, 0.9986, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5375, 0.5118, 0.5182, 0.5166, 0.5170, 0.5169,
```

```
0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5898])
finalReturns:  tensor([4.4805, 4.5534, 4.5271, 4.4326, 4.2901, 4.1135, 3.9125,
3.6938, 3.4621,
        3.2208, 2.9725, 2.7189, 2.4615, 2.2012, 1.9387, 1.6746, 1.4093, 1.1430,
        0.8761, 0.6087, 0.3410])
----------------------------------------
iter  0  stage  3  ep  0   adversary:  AdversaryModes.imitation_128
  actions:  tensor([27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,
27, 27, 27,
        27, 27, 27, 27, 27, 27,  0])
loss=  tensor(0.0272, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.6975,
6.6975, 6.6975, 6.6975, 6.1064, 5.6160, 5.1936, 4.8192, 4.4788,
        4.1629, 3.8647, 3.5795, 3.3038, 3.0352, 2.7719, 2.5124, 2.2558, 2.0013,
        1.7485, 1.4969, 1.2462, 0.9962, 0.7466, 0.4975, 0.2486]) return=
129338.69866666665
probs of actions:  tensor([0.9989, 0.9985, 0.9991, 0.9995, 0.9990, 0.9993,
0.9996, 0.9991, 1.0000,
        0.9994, 0.9995, 0.9995, 0.9998, 0.9998, 1.0000, 1.0000, 0.9997, 1.0000,
        1.0000, 1.0000, 1.0000, 0.9997, 0.9987, 0.9986, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5375, 0.5118, 0.5182, 0.5166, 0.5170, 0.5169,
0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5898])
finalReturns:  tensor([4.7487, 4.8216, 4.7954, 4.7008, 4.5583, 4.3818, 4.1808,
3.9621, 3.7303,
        3.4891, 3.2408, 2.9872, 2.7298, 2.4694, 2.2070, 1.9429, 1.6775, 1.4113,
        1.1444, 0.8770, 0.6092, 0.3412])
----------------------------------------
iter  0  stage  2  ep  0   adversary:  AdversaryModes.imitation_128
  actions:  tensor([27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,
27, 27, 27,
        27, 27, 27, 27, 27, 27,  0])
loss=  tensor(0.0326, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.9408,
6.9408, 6.9408, 6.3561, 5.8642, 5.4422, 5.0677, 4.7273, 4.4114,
        4.1132, 3.8280, 3.5523, 3.2837, 3.0203, 2.7608, 2.5042, 2.2498, 1.9970,
        1.7454, 1.4947, 1.2447, 0.9951, 0.7460, 0.4971, 0.2485]) return=
129338.69866666665
probs of actions:  tensor([0.9989, 0.9985, 0.9991, 0.9995, 0.9990, 0.9993,
0.9996, 0.9991, 1.0000,
        0.9994, 0.9995, 0.9995, 0.9998, 0.9998, 1.0000, 1.0000, 0.9997, 1.0000,
        1.0000, 1.0000, 1.0000, 0.9997, 0.9987, 0.9986, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5375, 0.5118, 0.5182, 0.5166, 0.5170, 0.5169,
0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169,
```

```
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5898])
finalReturns:  tensor([5.0173, 5.0902, 5.0638, 4.9693, 4.8268, 4.6503, 4.4493,
4.2305, 3.9988,
        3.7575, 3.5092, 3.2557, 2.9982, 2.7379, 2.4754, 2.2113, 1.9460, 1.6798,
        1.4129, 1.1455, 0.8777, 0.6096, 0.3413])
---------------------------------------
iter  0  stage  1  ep  594   adversary:  AdversaryModes.imitation_128
  actions:  tensor([27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,
27, 27, 27,
        27, 27, 27, 27, 27, 27,  0])
loss=  tensor(0.0351, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.2101,
7.2101, 6.5997, 6.1137, 5.6903, 5.3161, 4.9756, 4.6597, 4.3615,
        4.0763, 3.8007, 3.5321, 3.2687, 3.0092, 2.7526, 2.4982, 2.2454, 1.9938,
        1.7430, 1.4930, 1.2435, 0.9943, 0.7455, 0.4968, 0.2484]) return=
129338.69866666665
probs of actions:  tensor([0.9989, 0.9990, 0.9994, 0.9997, 0.9990, 0.9994,
0.9996, 0.9991, 1.0000,
        0.9994, 0.9995, 0.9995, 0.9998, 0.9998, 1.0000, 1.0000, 0.9997, 1.0000,
        1.0000, 1.0000, 1.0000, 0.9997, 0.9987, 0.9986, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5375, 0.5118, 0.5182, 0.5166, 0.5170, 0.5169,
0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5898])
finalReturns:  tensor([5.2854, 5.3583, 5.3326, 5.2378, 5.0954, 4.9188, 4.7178,
4.4991, 4.2673,
        4.0261, 3.7778, 3.5242, 3.2668, 3.0064, 2.7440, 2.4799, 2.2145, 1.9483,
        1.6814, 1.4140, 1.1462, 0.8782, 0.6099, 0.3415])
---------------------------------------
iter  0  stage  0  ep  69   adversary:  AdversaryModes.imitation_128
  actions:  tensor([27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,
27, 27, 27,
        27, 27, 27, 27, 27, 27,  0])
loss=  tensor(0.0403, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.3783,
6.8670, 6.3576, 5.9396, 5.5641, 5.2240, 4.9080, 4.6098, 4.3246,
        4.0490, 3.7804, 3.5170, 3.2575, 3.0009, 2.7465, 2.4936, 2.2420, 1.9913,
        1.7413, 1.4918, 1.2426, 0.9938, 0.7451, 0.4966, 0.2483]) return=
129338.69866666665
probs of actions:  tensor([0.9990, 0.9990, 0.9994, 0.9997, 0.9991, 0.9994,
0.9997, 0.9991, 1.0000,
        0.9994, 0.9995, 0.9995, 0.9998, 0.9998, 1.0000, 1.0000, 0.9997, 1.0000,
        1.0000, 1.0000, 1.0000, 0.9997, 0.9987, 0.9986, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5375, 0.5118, 0.5182, 0.5166, 0.5170, 0.5169,
0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5898])
finalReturns:  tensor([5.5556, 5.6285, 5.6005, 5.5067, 5.3639, 5.1875, 4.9865,
```

```
4.7677, 4.5360,
        4.2947, 4.0464, 3.7929, 3.5354, 3.2751, 3.0126, 2.7485, 2.4832, 2.2170,
        1.9501, 1.6827, 1.4149, 1.1468, 0.8786, 0.6101, 0.3416])
0,[1e-05,1][1, 10000, 1, 1],1682835632 saved
[1662531, 'tensor([0., 0., 0., 0., 1., 0., 0., 0., 0., 0., 0., 0.])',
129338.69866666665, 75332.17066666666, 0.04029531031847, 1e-05, 1, 0,
'tensor([27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,
27,\n        27, 27, 27, 27, 27, 27,  0])', '[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.\n 1.]', '0,[1e-05,1][1, 10000, 1,
1],1682835632', 25, 50, 172677.5835126837, 219861.0138117109, 91603.02777699233,
132330.31466666664, 129338.69866666666, 72150.91909080048, 72150.91909080048,
88030.91902269641, 88030.91902269641, 105848.51071772553, 72121.02259730823,
87961.7660416823]
policy reset
----------------------------------------
iter  1  stage  24  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([2, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
0, 4, 0, 0,
        0])
loss=  tensor(-0., grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.5107,
0.5107, 0.5107, 0.5107, 0.5107, 0.5107, 0.5107, 0.5107, 0.5107,
        0.5107, 0.5107, 0.5107, 0.5107, 0.5107, 0.5107, 0.5107, 0.5107, 0.5107,
        0.5107, 0.5107, 0.5107, 0.5107, 0.5107, 0.5107, 0.5107]) return=
127695.93011642047
probs of actions:  tensor([0.0087, 0.9502, 0.0086, 0.9312, 0.9305, 0.9451,
0.9223, 0.9450, 0.9302,
        0.9339, 0.9603, 0.9469, 0.9478, 0.9461, 0.9511, 0.9430, 0.0276, 0.9222,
        0.9427, 0.9465, 0.9359, 0.0034, 0.9471, 0.9232, 0.9943],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5108, 0.5166, 0.5077, 0.5174, 0.5079, 0.5103, 0.5097,
0.5098, 0.5098,
        0.5098, 0.5098, 0.5098, 0.5098, 0.5098, 0.5098, 0.5098, 0.5097, 0.5134,
        0.5089, 0.5100, 0.5097, 0.5082, 0.5242, 0.5062, 0.5107])
finalReturns:  tensor([0.])
----------------------------------------
iter  1  stage  23  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([18, 18, 18,  0, 18,  6,  0,  0,  1,  0,  0,  0,  0, 18,  0,
0,  0, 18,
        18,  0,  0, 18,  0, 18,  0])
loss=  tensor(0.0075, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.0078,
1.0078, 1.0078, 1.0078, 1.0078, 1.0078, 1.0078, 1.0078, 1.0078,
        1.0078, 1.0078, 1.0078, 1.0078, 1.0078, 1.0078, 1.0078, 1.0078, 1.0078,
        1.0078, 1.0078, 1.0078, 1.0078, 1.0078, 1.0078, 0.5138]) return=
129636.67739657205
probs of actions:  tensor([0.3905, 0.2322, 0.2606, 0.3254, 0.3457, 0.0249,
0.3994, 0.4527, 0.0195,
        0.5938, 0.7716, 0.4525, 0.6561, 0.3951, 0.5880, 0.6364, 0.4128, 0.3614,
        0.4197, 0.4487, 0.4621, 0.3929, 0.5632, 0.8044, 0.9996],
```

```
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5433, 0.5268, 0.5633, 0.4644, 0.5760, 0.5143,
0.5087, 0.5100,
        0.5133, 0.5089, 0.5100, 0.5097, 0.4774, 0.5761, 0.4939, 0.5138, 0.4764,
        0.5439, 0.5591, 0.4978, 0.4804, 0.5753, 0.4616, 0.5803])
finalReturns:  tensor([0.0341, 0.0665])
----------------------------------------
iter  1  stage  22  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        16, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0044, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.5102,
1.5102, 1.5102, 1.5102, 1.5102, 1.5102, 1.5102, 1.5102, 1.5102,
        1.5102, 1.5102, 1.5102, 1.5102, 1.5102, 1.5102, 1.5102, 1.5102, 1.5102,
        1.5102, 1.5102, 1.5102, 1.5102, 1.5102, 0.9476, 0.4506]) return=
132450.34798175492
probs of actions:  tensor([0.9696, 0.9712, 0.9630, 0.9755, 0.9730, 0.9697,
0.9775, 0.9711, 0.9708,
        0.9457, 0.9814, 0.9735, 0.9778, 0.9767, 0.9721, 0.9678, 0.9664, 0.9676,
        0.0071, 0.9770, 0.9657, 0.9756, 0.9838, 0.9828, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5433, 0.5268, 0.5309, 0.5299, 0.5302, 0.5301,
0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301,
        0.5369, 0.5226, 0.5320, 0.5296, 0.5302, 0.5301, 0.5625])
finalReturns:  tensor([0.1126, 0.1450, 0.1119])
----------------------------------------
iter  1  stage  21  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0026, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.9273,
1.9273, 1.9273, 1.9273, 1.9273, 1.9273, 1.9273, 1.9273, 1.9273,
        1.9273, 1.9273, 1.9273, 1.9273, 1.9273, 1.9273, 1.9273, 1.9273, 1.9273,
        1.9273, 1.9273, 1.9273, 1.9273, 1.3648, 0.8678, 0.4172]) return=
132442.06666666665
probs of actions:  tensor([0.9932, 0.9936, 0.9917, 0.9944, 0.9938, 0.9933,
0.9947, 0.9937, 0.9934,
        0.9874, 0.9960, 0.9942, 0.9954, 0.9948, 0.9939, 0.9928, 0.9918, 0.9923,
        0.9933, 0.9947, 0.9919, 0.9962, 0.9969, 0.9958, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5433, 0.5268, 0.5309, 0.5299, 0.5302, 0.5301,
0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5625])
finalReturns:  tensor([0.2255, 0.2579, 0.2248, 0.1453])
----------------------------------------
iter  1  stage  20  ep  81988   adversary:  AdversaryModes.imitation_128
```

```
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0010, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.3204,
2.3204, 2.3204, 2.3204, 2.3204, 2.3204, 2.3204, 2.3204, 2.3204,
        2.3204, 2.3204, 2.3204, 2.3204, 2.3204, 2.3204, 2.3204, 2.3204, 2.3204,
        2.3204, 2.3204, 2.3204, 1.7579, 1.2609, 0.8103, 0.3931]) return=
132442.06666666665
probs of actions:  tensor([0.9987, 0.9987, 0.9984, 0.9988, 0.9987, 0.9986,
0.9989, 0.9988, 0.9986,
        0.9973, 0.9993, 0.9988, 0.9991, 0.9989, 0.9988, 0.9986, 0.9981, 0.9983,
        0.9986, 0.9989, 0.9990, 0.9992, 0.9999, 0.9990, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5433, 0.5268, 0.5309, 0.5299, 0.5302, 0.5301,
0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5625])
finalReturns:  tensor([0.3625, 0.3949, 0.3618, 0.2823, 0.1694])
----------------------------------------
iter  1  stage  19  ep  56   adversary:  AdversaryModes.imitation_128
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0018, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.6958,
2.6958, 2.6958, 2.6958, 2.6958, 2.6958, 2.6958, 2.6958, 2.6958,
        2.6958, 2.6958, 2.6958, 2.6958, 2.6958, 2.6958, 2.6958, 2.6958, 2.6958,
        2.6958, 2.6958, 2.1333, 1.6363, 1.1857, 0.7685, 0.3754]) return=
132442.06666666665
probs of actions:  tensor([0.9986, 0.9987, 0.9983, 0.9988, 0.9986, 0.9985,
0.9988, 0.9987, 0.9986,
        0.9973, 0.9992, 0.9988, 0.9991, 0.9989, 0.9988, 0.9986, 0.9981, 0.9983,
        0.9985, 0.9990, 0.9990, 0.9992, 0.9999, 0.9989, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5433, 0.5268, 0.5309, 0.5299, 0.5302, 0.5301,
0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5625])
finalReturns:  tensor([0.5172, 0.5496, 0.5165, 0.4370, 0.3241, 0.1871])
----------------------------------------
iter  1  stage  18  ep  344   adversary:  AdversaryModes.imitation_128
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0027, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.0583,
3.0583, 3.0583, 3.0583, 3.0583, 3.0583, 3.0583, 3.0583, 3.0583,
        3.0583, 3.0583, 3.0583, 3.0583, 3.0583, 3.0583, 3.0583, 3.0583, 3.0583,
        3.0583, 2.4958, 1.9987, 1.5482, 1.1309, 0.7379, 0.3624]) return=
132442.06666666665
```

```
probs of actions:  tensor([0.9987, 0.9987, 0.9984, 0.9988, 0.9987, 0.9986,
0.9989, 0.9988, 0.9986,
        0.9974, 0.9993, 0.9988, 0.9992, 0.9989, 0.9988, 0.9987, 0.9982, 0.9984,
        0.9990, 0.9993, 0.9990, 0.9993, 0.9999, 0.9990, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5433, 0.5268, 0.5309, 0.5299, 0.5302, 0.5301,
0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5625])
finalReturns:  tensor([0.6848, 0.7172, 0.6842, 0.6046, 0.4918, 0.3547, 0.2001])
----------------------------------------
iter  1  stage  17  ep  653   adversary:  AdversaryModes.imitation_128
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0039, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.4111,
3.4111, 3.4111, 3.4111, 3.4111, 3.4111, 3.4111, 3.4111, 3.4111,
        3.4111, 3.4111, 3.4111, 3.4111, 3.4111, 3.4111, 3.4111, 3.4111, 3.4111,
        2.8486, 2.3516, 1.9010, 1.4838, 1.0907, 0.7153, 0.3529]) return=
132442.06666666665
probs of actions:  tensor([0.9987, 0.9988, 0.9985, 0.9988, 0.9987, 0.9986,
0.9989, 0.9988, 0.9986,
        0.9974, 0.9993, 0.9988, 0.9992, 0.9989, 0.9988, 0.9987, 0.9982, 0.9990,
        0.9991, 0.9994, 0.9990, 0.9993, 0.9999, 0.9990, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5433, 0.5268, 0.5309, 0.5299, 0.5302, 0.5301,
0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5625])
finalReturns:  tensor([0.8621, 0.8945, 0.8614, 0.7819, 0.6690, 0.5320, 0.3773,
0.2096])
----------------------------------------
iter  1  stage  16  ep  23347   adversary:  AdversaryModes.imitation_128
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0031, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.7569,
3.7569, 3.7569, 3.7569, 3.7569, 3.7569, 3.7569, 3.7569, 3.7569,
        3.7569, 3.7569, 3.7569, 3.7569, 3.7569, 3.7569, 3.7569, 3.7569, 3.1944,
        2.6974, 2.2468, 1.8296, 1.4365, 1.0611, 0.6986, 0.3458]) return=
132442.06666666665
probs of actions:  tensor([0.9992, 0.9993, 0.9991, 0.9993, 0.9992, 0.9991,
0.9993, 0.9993, 0.9992,
        0.9985, 0.9996, 0.9993, 0.9995, 0.9994, 0.9993, 0.9992, 0.9990, 0.9996,
        0.9995, 0.9997, 0.9995, 0.9998, 1.0000, 0.9996, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5433, 0.5268, 0.5309, 0.5299, 0.5302, 0.5301,
0.5301, 0.5301,
```

```
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5625])
finalReturns:  tensor([1.0464, 1.0788, 1.0457, 0.9662, 0.8533, 0.7163, 0.5616,
0.3940, 0.2167])
----------------------------------------
iter  1  stage  15  ep  0   adversary:  AdversaryModes.imitation_128
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0044, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.0974,
4.0974, 4.0974, 4.0974, 4.0974, 4.0974, 4.0974, 4.0974, 4.0974,
        4.0974, 4.0974, 4.0974, 4.0974, 4.0974, 4.0974, 4.0974, 3.5349, 3.0379,
        2.5873, 2.1701, 1.7770, 1.4016, 1.0391, 0.6863, 0.3405]) return=
132442.06666666665
probs of actions:  tensor([0.9992, 0.9993, 0.9991, 0.9993, 0.9992, 0.9991,
0.9993, 0.9993, 0.9992,
        0.9985, 0.9996, 0.9993, 0.9995, 0.9994, 0.9993, 0.9992, 0.9990, 0.9996,
        0.9995, 0.9997, 0.9995, 0.9998, 1.0000, 0.9996, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5433, 0.5268, 0.5309, 0.5299, 0.5302, 0.5301,
0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5625])
finalReturns:  tensor([1.2360, 1.2684, 1.2353, 1.1558, 1.0429, 0.9059, 0.7512,
0.5836, 0.4063,
        0.2220])
----------------------------------------
iter  1  stage  14  ep  0   adversary:  AdversaryModes.imitation_128
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0059, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.4340,
4.4340, 4.4340, 4.4340, 4.4340, 4.4340, 4.4340, 4.4340, 4.4340,
        4.4340, 4.4340, 4.4340, 4.4340, 4.4340, 4.4340, 4.4340, 3.8715, 3.3744, 2.9239,
        2.5066, 2.1136, 1.7381, 1.3757, 1.0228, 0.6770, 0.3366]) return=
132442.06666666665
probs of actions:  tensor([0.9992, 0.9993, 0.9991, 0.9993, 0.9992, 0.9991,
0.9993, 0.9993, 0.9992,
        0.9985, 0.9996, 0.9993, 0.9995, 0.9994, 0.9993, 0.9992, 0.9990, 0.9996,
        0.9995, 0.9997, 0.9995, 0.9998, 1.0000, 0.9996, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5433, 0.5268, 0.5309, 0.5299, 0.5302, 0.5301,
0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5625])
finalReturns:  tensor([1.4295, 1.4619, 1.4289, 1.3493, 1.2365, 1.0994, 0.9448,
0.7771, 0.5999,
        0.4156, 0.2259])
```

```
----------------------------------------
iter  1  stage  13  ep  0   adversary:  AdversaryModes.imitation_128
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0074, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.7676,
4.7676, 4.7676, 4.7676, 4.7676, 4.7676, 4.7676, 4.7676, 4.7676,
        4.7676, 4.7676, 4.7676, 4.7676, 4.7676, 4.2051, 3.7081, 3.2575, 2.8402,
        2.4472, 2.0718, 1.7093, 1.3564, 1.0107, 0.6702, 0.3336]) return=
132442.06666666665
probs of actions:  tensor([0.9992, 0.9993, 0.9991, 0.9993, 0.9992, 0.9991,
0.9993, 0.9993, 0.9992,
        0.9985, 0.9996, 0.9993, 0.9995, 0.9994, 0.9993, 0.9992, 0.9990, 0.9996,
        0.9995, 0.9997, 0.9995, 0.9998, 1.0000, 0.9996, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5433, 0.5268, 0.5309, 0.5299, 0.5302, 0.5301,
0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5625])
finalReturns:  tensor([1.6260, 1.6584, 1.6253, 1.5458, 1.4330, 1.2959, 1.1412,
0.9736, 0.7964,
        0.6120, 0.4224, 0.2289])
----------------------------------------
iter  1  stage  12  ep  0   adversary:  AdversaryModes.imitation_128
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0088, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.0990,
5.0990, 5.0990, 5.0990, 5.0990, 5.0990, 5.0990, 5.0990, 5.0990,
        5.0990, 5.0990, 5.0990, 5.0990, 4.5365, 4.0395, 3.5889, 3.1717, 2.7786,
        2.4032, 2.0407, 1.6879, 1.3421, 1.0016, 0.6651, 0.3314]) return=
132442.06666666665
probs of actions:  tensor([0.9992, 0.9993, 0.9991, 0.9993, 0.9992, 0.9991,
0.9993, 0.9993, 0.9992,
        0.9985, 0.9996, 0.9993, 0.9995, 0.9994, 0.9993, 0.9992, 0.9990, 0.9996,
        0.9995, 0.9997, 0.9995, 0.9998, 1.0000, 0.9996, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5433, 0.5268, 0.5309, 0.5299, 0.5302, 0.5301,
0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5625])
finalReturns:  tensor([1.8247, 1.8571, 1.8240, 1.7445, 1.6316, 1.4946, 1.3399,
1.1723, 0.9950,
        0.8107, 0.6211, 0.4275, 0.2311])
----------------------------------------
iter  1  stage  11  ep  0   adversary:  AdversaryModes.imitation_128
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
```

```
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0107, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.4288,
5.4288, 5.4288, 5.4288, 5.4288, 5.4288, 5.4288, 5.4288, 5.4288,
        5.4288, 5.4288, 5.4288, 4.8663, 4.3693, 3.9187, 3.5015, 3.1084, 2.7330,
        2.3705, 2.0177, 1.6719, 1.3314, 0.9949, 0.6612, 0.3298]) return=
132442.06666666665
probs of actions:  tensor([0.9992, 0.9993, 0.9991, 0.9993, 0.9992, 0.9991,
0.9993, 0.9993, 0.9992,
        0.9985, 0.9996, 0.9993, 0.9995, 0.9994, 0.9993, 0.9992, 0.9990, 0.9996,
        0.9995, 0.9997, 0.9995, 0.9998, 1.0000, 0.9996, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5433, 0.5268, 0.5309, 0.5299, 0.5302, 0.5301,
0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5625])
finalReturns:  tensor([2.0250, 2.0574, 2.0243, 1.9448, 1.8319, 1.6949, 1.5402,
1.3726, 1.1953,
        1.0110, 0.8214, 0.6278, 0.4314, 0.2327])
----------------------------------------
iter  1  stage  10  ep  0   adversary:  AdversaryModes.imitation_128
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0121, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.7574,
5.7574, 5.7574, 5.7574, 5.7574, 5.7574, 5.7574, 5.7574, 5.7574,
        5.7574, 5.7574, 5.1949, 4.6979, 4.2473, 3.8300, 3.4370, 3.0615, 2.6991,
        2.3462, 2.0005, 1.6600, 1.3234, 0.9898, 0.6584, 0.3286]) return=
132442.06666666665
probs of actions:  tensor([0.9992, 0.9993, 0.9991, 0.9993, 0.9992, 0.9991,
0.9993, 0.9993, 0.9992,
        0.9985, 0.9996, 0.9993, 0.9995, 0.9994, 0.9993, 0.9992, 0.9990, 0.9996,
        0.9995, 0.9997, 0.9995, 0.9998, 1.0000, 0.9996, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5433, 0.5268, 0.5309, 0.5299, 0.5302, 0.5301,
0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5625])
finalReturns:  tensor([2.2265, 2.2589, 2.2258, 2.1463, 2.0335, 1.8964, 1.7418,
1.5741, 1.3969,
        1.2125, 1.0229, 0.8294, 0.6329, 0.4342, 0.2339])
----------------------------------------
iter  1  stage  9  ep  207   adversary:  AdversaryModes.imitation_128
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0116, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.0850,
6.0850, 6.0850, 6.0850, 6.0850, 6.0850, 6.0850, 6.0850, 6.0850,
        6.0850, 5.5225, 5.0255, 4.5749, 4.1577, 3.7646, 3.3892, 3.0267, 2.6739,
```

```
        2.3281, 1.9876, 1.6511, 1.3174, 0.9860, 0.6562, 0.3276]) return=
132442.06666666665
probs of actions:  tensor([0.9993, 0.9994, 0.9992, 0.9994, 0.9993, 0.9993,
0.9994, 0.9994, 0.9993,
        0.9990, 0.9997, 0.9995, 0.9997, 0.9996, 0.9995, 0.9995, 0.9991, 0.9996,
        0.9996, 0.9997, 0.9996, 0.9998, 1.0000, 0.9997, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5433, 0.5268, 0.5309, 0.5299, 0.5302, 0.5301,
0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5625])
finalReturns:  tensor([2.4290, 2.4614, 2.4283, 2.3488, 2.2359, 2.0989, 1.9442,
1.7766, 1.5993,
        1.4150, 1.2254, 1.0318, 0.8354, 0.6367, 0.4364, 0.2349])
----------------------------------------
iter  1  stage  8  ep  0   adversary:  AdversaryModes.imitation_128
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0139, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.4120,
6.4120, 6.4120, 6.4120, 6.4120, 6.4120, 6.4120, 6.4120, 6.4120,
        5.8495, 5.3525, 4.9019, 4.4846, 4.0916, 3.7162, 3.3537, 3.0008, 2.6551,
        2.3146, 1.9780, 1.6444, 1.3130, 0.9832, 0.6546, 0.3270]) return=
132442.06666666665
probs of actions:  tensor([0.9993, 0.9994, 0.9992, 0.9994, 0.9993, 0.9993,
0.9994, 0.9994, 0.9993,
        0.9990, 0.9997, 0.9995, 0.9997, 0.9996, 0.9995, 0.9995, 0.9991, 0.9996,
        0.9996, 0.9997, 0.9996, 0.9998, 1.0000, 0.9997, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5433, 0.5268, 0.5309, 0.5299, 0.5302, 0.5301,
0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5625])
finalReturns:  tensor([2.6321, 2.6645, 2.6314, 2.5519, 2.4391, 2.3020, 2.1473,
1.9797, 1.8025,
        1.6181, 1.4285, 1.2350, 1.0385, 0.8398, 0.6395, 0.4380, 0.2355])
----------------------------------------
iter  1  stage  7  ep  0   adversary:  AdversaryModes.imitation_128
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0161, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.7384,
6.7384, 6.7384, 6.7384, 6.7384, 6.7384, 6.7384, 6.7384, 6.1759,
        5.6789, 5.2283, 4.8111, 4.4180, 4.0426, 3.6802, 3.3273, 2.9815, 2.6410,
        2.3045, 1.9708, 1.6394, 1.3096, 0.9811, 0.6534, 0.3264]) return=
132442.06666666665
probs of actions:  tensor([0.9993, 0.9994, 0.9992, 0.9994, 0.9993, 0.9993,
0.9994, 0.9994, 0.9993,
```

```
        0.9990, 0.9997, 0.9995, 0.9997, 0.9996, 0.9995, 0.9995, 0.9991, 0.9996,
        0.9996, 0.9997, 0.9996, 0.9998, 1.0000, 0.9997, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5433, 0.5268, 0.5309, 0.5299, 0.5302, 0.5301,
0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5625])
finalReturns:  tensor([2.8358, 2.8682, 2.8351, 2.7556, 2.6427, 2.5057, 2.3510,
2.1833, 2.0061,
        1.8218, 1.6322, 1.4386, 1.2422, 1.0435, 0.8432, 0.6416, 0.4392, 0.2361])
---------------------------------------
iter  1  stage  6  ep  0   adversary:  AdversaryModes.imitation_128
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0183, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.0645,
7.0645, 7.0645, 7.0645, 7.0645, 7.0645, 7.0645, 6.5020, 6.0050,
        5.5544, 5.1371, 4.7441, 4.3687, 4.0062, 3.6533, 3.3076, 2.9671, 2.6305,
        2.2969, 1.9655, 1.6357, 1.3071, 0.9795, 0.6525, 0.3261]) return=
132442.06666666665
probs of actions:  tensor([0.9993, 0.9994, 0.9992, 0.9994, 0.9993, 0.9993,
0.9994, 0.9994, 0.9993,
        0.9990, 0.9997, 0.9995, 0.9997, 0.9996, 0.9995, 0.9995, 0.9991, 0.9996,
        0.9996, 0.9997, 0.9996, 0.9998, 1.0000, 0.9997, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5433, 0.5268, 0.5309, 0.5299, 0.5302, 0.5301,
0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5625])
finalReturns:  tensor([3.0398, 3.0722, 3.0391, 2.9596, 2.8467, 2.7097, 2.5550,
2.3874, 2.2102,
        2.0258, 1.8362, 1.6427, 1.4462, 1.2475, 1.0472, 0.8457, 0.6432, 0.4401,
        0.2364])
---------------------------------------
iter  1  stage  5  ep  0   adversary:  AdversaryModes.imitation_128
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0212, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.3903,
7.3903, 7.3903, 7.3903, 7.3903, 7.3903, 6.8277, 6.3307, 5.8802,
        5.4629, 5.0698, 4.6944, 4.3320, 3.9791, 3.6333, 3.2929, 2.9563, 2.6227,
        2.2912, 1.9614, 1.6329, 1.3052, 0.9783, 0.6518, 0.3258]) return=
132442.06666666665
probs of actions:  tensor([0.9993, 0.9994, 0.9992, 0.9994, 0.9993, 0.9993,
0.9994, 0.9994, 0.9993,
        0.9990, 0.9997, 0.9995, 0.9997, 0.9996, 0.9995, 0.9995, 0.9991, 0.9996,
        0.9996, 0.9997, 0.9996, 0.9998, 1.0000, 0.9997, 1.0000],
       grad_fn=<ExpBackward0>)
```

```
rewards:  tensor([0.4788, 0.5433, 0.5268, 0.5309, 0.5299, 0.5302, 0.5301,
0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5625])
finalReturns:  tensor([3.2441, 3.2765, 3.2435, 3.1639, 3.0511, 2.9141, 2.7594,
2.5917, 2.4145,
        2.2302, 2.0405, 1.8470, 1.6505, 1.4519, 1.2516, 1.0500, 0.8476, 0.6444,
        0.4408, 0.2367])
----------------------------------------
iter  1  stage  4  ep  0   adversary:  AdversaryModes.imitation_128
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0241, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.7156,
7.7156, 7.7156, 7.7156, 7.7156, 7.1533, 6.6563, 6.2057, 5.7885,
        5.3954, 5.0200, 4.6575, 4.3047, 3.9589, 3.6184, 3.2818, 2.9482, 2.6168,
        2.2870, 1.9584, 1.6308, 1.3038, 0.9774, 0.6513, 0.3256]) return=
132442.06666666665
probs of actions:  tensor([0.9993, 0.9994, 0.9992, 0.9994, 0.9993, 0.9993,
0.9994, 0.9994, 0.9993,
        0.9990, 0.9997, 0.9995, 0.9997, 0.9996, 0.9995, 0.9995, 0.9991, 0.9996,
        0.9996, 0.9997, 0.9996, 0.9998, 1.0000, 0.9997, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5433, 0.5268, 0.5309, 0.5299, 0.5302, 0.5301,
0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5625])
finalReturns:  tensor([3.4487, 3.4811, 3.4480, 3.3685, 3.2556, 3.1186, 2.9639,
2.7963, 2.6190,
        2.4347, 2.2451, 2.0516, 1.8551, 1.6564, 1.4561, 1.2546, 1.0521, 0.8490,
        0.6453, 0.4413, 0.2369])
----------------------------------------
iter  1  stage  3  ep  0   adversary:  AdversaryModes.imitation_128
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0268, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.0419,
8.0419, 8.0419, 8.0419, 7.4785, 6.9817, 6.5311, 6.1139, 5.7208,
        5.3454, 4.9829, 4.6301, 4.2843, 3.9438, 3.6072, 3.2736, 2.9422, 2.6124,
        2.2838, 1.9562, 1.6292, 1.3028, 0.9767, 0.6509, 0.3254]) return=
132442.06666666665
probs of actions:  tensor([0.9993, 0.9994, 0.9992, 0.9994, 0.9993, 0.9993,
0.9994, 0.9994, 0.9993,
        0.9990, 0.9997, 0.9995, 0.9997, 0.9996, 0.9995, 0.9995, 0.9991, 0.9996,
        0.9996, 0.9997, 0.9996, 0.9998, 1.0000, 0.9997, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5433, 0.5268, 0.5309, 0.5299, 0.5302, 0.5301,
0.5301, 0.5301,
```

```
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5625])
finalReturns:  tensor([3.6534, 3.6858, 3.6527, 3.5732, 3.4603, 3.3233, 3.1686,
3.0010, 2.8237,
        2.6394, 2.4498, 2.2563, 2.0598, 1.8611, 1.6608, 1.4593, 1.2568, 1.0537,
        0.8500, 0.6460, 0.4417, 0.2371])
----------------------------------------
iter  1  stage  2  ep  0   adversary:  AdversaryModes.imitation_128
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0304, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.3638,
8.3638, 8.3638, 7.8046, 7.3068, 6.8564, 6.4391, 6.0461, 5.6706,
        5.3082, 4.9553, 4.6095, 4.2691, 3.9325, 3.5989, 3.2674, 2.9377, 2.6091,
        2.2814, 1.9545, 1.6280, 1.3020, 0.9762, 0.6507, 0.3253]) return=
132442.06666666665
probs of actions:  tensor([0.9993, 0.9994, 0.9992, 0.9994, 0.9993, 0.9993,
0.9994, 0.9994, 0.9993,
        0.9990, 0.9997, 0.9995, 0.9997, 0.9996, 0.9995, 0.9995, 0.9991, 0.9996,
        0.9996, 0.9997, 0.9996, 0.9998, 1.0000, 0.9997, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5433, 0.5268, 0.5309, 0.5299, 0.5302, 0.5301,
0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5625])
finalReturns:  tensor([3.8583, 3.8907, 3.8575, 3.7780, 3.6652, 3.5281, 3.3735,
3.2058, 3.0286,
        2.8442, 2.6546, 2.4611, 2.2646, 2.0660, 1.8656, 1.6641, 1.4617, 1.2585,
        1.0549, 0.8508, 0.6465, 0.4419, 0.2372])
----------------------------------------
iter  1  stage  1  ep  0   adversary:  AdversaryModes.imitation_128
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0335, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.7023,
8.7023, 8.1266, 7.6327, 7.1814, 6.7643, 6.3712, 5.9958, 5.6334,
        5.2805, 4.9347, 4.5942, 4.2577, 3.9241, 3.5926, 3.2628, 2.9343, 2.6066,
        2.2797, 1.9532, 1.6271, 1.3014, 0.9758, 0.6504, 0.3252]) return=
132442.06666666665
probs of actions:  tensor([0.9993, 0.9994, 0.9992, 0.9994, 0.9993, 0.9993,
0.9994, 0.9994, 0.9993,
        0.9990, 0.9997, 0.9995, 0.9997, 0.9996, 0.9995, 0.9995, 0.9991, 0.9996,
        0.9996, 0.9997, 0.9996, 0.9998, 1.0000, 0.9997, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5433, 0.5268, 0.5309, 0.5299, 0.5302, 0.5301,
0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5625])
```

```
finalReturns:  tensor([4.0630, 4.0954, 4.0626, 3.9829, 3.8701, 3.7331, 3.5784,
3.4107, 3.2335,
        3.0492, 2.8596, 2.6660, 2.4695, 2.2709, 2.0706, 1.8690, 1.6666, 1.4634,
        1.2598, 1.0558, 0.8514, 0.6469, 0.4422, 0.2373])
---------------------------------------
iter  1  stage  0  ep  0   adversary:  AdversaryModes.imitation_128
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0369, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.9755,
8.4643, 7.9549, 7.5072, 7.0893, 6.6964, 6.3209, 5.9585, 5.6056,
        5.2598, 4.9193, 4.5828, 4.2492, 3.9177, 3.5879, 3.2594, 2.9317, 2.6048,
        2.2783, 1.9523, 1.6265, 1.3009, 0.9755, 0.6503, 0.3251]) return=
132442.06666666665
probs of actions:  tensor([0.9993, 0.9994, 0.9992, 0.9994, 0.9993, 0.9993,
0.9994, 0.9994, 0.9993,
        0.9990, 0.9997, 0.9996, 0.9997, 0.9996, 0.9996, 0.9995, 0.9991, 0.9996,
        0.9996, 0.9997, 0.9996, 0.9998, 1.0000, 0.9997, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5433, 0.5268, 0.5309, 0.5299, 0.5302, 0.5301,
0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301,
        0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5301, 0.5625])
finalReturns:  tensor([4.2687, 4.3011, 4.2672, 4.1880, 4.0751, 3.9381, 3.7834,
3.6157, 3.4385,
        3.2542, 3.0646, 2.8710, 2.6745, 2.4759, 2.2756, 2.0740, 1.8716, 1.6684,
        1.4648, 1.2607, 1.0564, 0.8519, 0.6472, 0.4423, 0.2374])
0,[1e-05,1][1, 10000, 1, 1],1682849652 saved
[586616, 'tensor([0., 0., 0., 0., 1., 0., 0., 0., 0., 0., 0., 0.])',
132442.06666666665, 85788.46666666665, 0.036905523389577866, 1e-05, 1, 0,
'tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18,\n        18, 18, 18, 18, 18, 18,  0])', '[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.\n 1.]', '0,[1e-05,1][1, 10000, 1,
1],1682849652', 25, 50, 164645.8985006173, 193521.60851536895,
78111.91688437786, 135363.21866666665, 132442.06666666665, 102604.52812037412,
102604.52812037412, 120699.79187222247, 120699.79187222247, 90646.446578478,
102630.27987904308, 120699.79187222247]
policy reset
---------------------------------------
iter  2  stage  24  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0,
0, 0, 0, 0,
        0])
loss=  tensor(-0., grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.5098,
0.5098, 0.5098, 0.5098, 0.5098, 0.5098, 0.5098, 0.5098, 0.5098,
        0.5098, 0.5098, 0.5098, 0.5098, 0.5098, 0.5098, 0.5098, 0.5098, 0.5098,
        0.5098, 0.5098, 0.5098, 0.5098, 0.5098, 0.5098, 0.5098]) return=
127541.45169466147
```

```
probs of actions:  tensor([0.8878, 0.8998, 0.9017, 0.0455, 0.8904, 0.8897,
0.8844, 0.9010, 0.9006,
        0.9017, 0.8732, 0.9087, 0.8745, 0.8932, 0.8736, 0.9308, 0.8922, 0.0170,
        0.8716, 0.8762, 0.8754, 0.8779, 0.8982, 0.9095, 0.9904],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5112, 0.5094, 0.5099, 0.5097, 0.5134, 0.5089, 0.5100,
0.5097, 0.5098,
        0.5098, 0.5098, 0.5098, 0.5098, 0.5098, 0.5098, 0.5098, 0.5098, 0.5094,
        0.5170, 0.5080, 0.5102, 0.5097, 0.5098, 0.5098, 0.5098])
finalReturns:  tensor([0.])
----------------------------------------
iter  2  stage  23  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([ 0, 16,  0, 10,  3, 17,  1, 19,  1,  0,  1, 19,  1,  0,  0,
0,  1,  1,
         0, 19,  0, 19,  0, 25,  0])
loss=  tensor(0.1094, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.0062,
1.0062, 1.0062, 1.0062, 1.0062, 1.0062, 1.0062, 1.0062, 1.0062,
        1.0062, 1.0062, 1.0062, 1.0062, 1.0062, 1.0062, 1.0062, 1.0062, 1.0062,
        1.0062, 1.0062, 1.0062, 1.0062, 1.0062, 1.0062, 0.5143]) return=
129474.82480479611
probs of actions:  tensor([0.3228, 0.0171, 0.2714, 0.0419, 0.0283, 0.0064,
0.0947, 0.2192, 0.0974,
        0.4085, 0.0975, 0.1958, 0.0921, 0.3377, 0.2460, 0.5109, 0.0855, 0.0916,
        0.2574, 0.2246, 0.2617, 0.2131, 0.3345, 0.0302, 0.9922],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5112, 0.4838, 0.5686, 0.4856, 0.5489, 0.4818, 0.5720,
0.4622, 0.5829,
        0.4958, 0.5132, 0.4764, 0.5791, 0.4967, 0.5131, 0.5090, 0.5099, 0.5132,
        0.5125, 0.4730, 0.5801, 0.4568, 0.5844, 0.4294, 0.6079])
finalReturns:  tensor([0.0311, 0.0936])
----------------------------------------
iter  2  stage  22  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([19, 19, 19, 19, 19, 20, 19, 19, 19, 19, 19, 19, 19, 19, 19,
19, 19, 19,
        19, 19, 19, 19, 19, 19,  0])
loss=  tensor(0.0571, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.5092,
1.5092, 1.5092, 1.5092, 1.5092, 1.5092, 1.5092, 1.5092, 1.5092,
        1.5092, 1.5092, 1.5092, 1.5092, 1.5092, 1.5092, 1.5092, 1.5092, 1.5092,
        1.5092, 1.5092, 1.5092, 1.5092, 1.5092, 0.9437, 0.4474]) return=
132272.56485156258
probs of actions:  tensor([0.8868, 0.8613, 0.8450, 0.8907, 0.8717, 0.0441,
0.8733, 0.8735, 0.9006,
        0.8514, 0.8717, 0.8657, 0.8498, 0.8412, 0.8213, 0.8708, 0.8853, 0.8783,
        0.8333, 0.8497, 0.8522, 0.8566, 0.9069, 0.7395, 0.9979],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4751, 0.5434, 0.5259, 0.5303, 0.5292, 0.5256, 0.5332,
0.5285, 0.5296,
        0.5293, 0.5294, 0.5294, 0.5294, 0.5294, 0.5294, 0.5294, 0.5294, 0.5294,
```

```
        0.5294, 0.5294, 0.5294, 0.5294, 0.5294, 0.5294, 0.5655])
finalReturns:  tensor([0.1151, 0.1512, 0.1181])
----------------------------------------
iter  2  stage  21  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19,
19, 19, 19,
        19, 19, 19, 19, 19, 20,  0])
loss=  tensor(0.5379, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.9216,
1.9216, 1.9216, 1.9216, 1.9216, 1.9216, 1.9216, 1.9216, 1.9216,
        1.9216, 1.9216, 1.9216, 1.9216, 1.9216, 1.9216, 1.9216, 1.9216, 1.9216,
        1.9216, 1.9216, 1.9216, 1.9216, 1.3561, 0.8598, 0.4124]) return=
132280.08116666667
probs of actions:  tensor([0.9403, 0.9196, 0.9068, 0.9415, 0.9396, 0.9242,
0.9294, 0.9377, 0.9528,
        0.9275, 0.9298, 0.9316, 0.9182, 0.9129, 0.8965, 0.9457, 0.9382, 0.9369,
        0.9005, 0.9167, 0.9171, 0.9177, 0.9507, 0.1169, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4751, 0.5434, 0.5259, 0.5303, 0.5292, 0.5295, 0.5294,
0.5294, 0.5294,
        0.5294, 0.5294, 0.5294, 0.5294, 0.5294, 0.5294, 0.5294, 0.5294, 0.5294,
        0.5294, 0.5294, 0.5294, 0.5294, 0.5294, 0.5255, 0.5693])
finalReturns:  tensor([0.2320, 0.2681, 0.2350, 0.1569])
----------------------------------------
iter  2  stage  20  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([19, 20, 19, 19, 19, 19, 19, 20, 19, 19, 19, 19, 19, 19, 19,
19, 20, 19,
        19, 19, 19, 20, 19, 19,  0])
loss=  tensor(1.0797, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.3038,
2.3038, 2.3038, 2.3038, 2.3038, 2.3038, 2.3038, 2.3038, 2.3038,
        2.3038, 2.3038, 2.3038, 2.3038, 2.3038, 2.3038, 2.3038, 2.3038, 2.3038,
        2.3038, 2.3038, 2.3038, 1.7384, 1.2420, 0.7946, 0.3855]) return=
132246.351396498
probs of actions:  tensor([0.8866, 0.1261, 0.8280, 0.8925, 0.8865, 0.8621,
0.8719, 0.0946, 0.9101,
        0.8700, 0.8727, 0.8724, 0.8557, 0.8425, 0.8268, 0.8876, 0.0950, 0.8847,
        0.8229, 0.8516, 0.8295, 0.1281, 0.9126, 0.6660, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4751, 0.5395, 0.5297, 0.5293, 0.5294, 0.5294, 0.5294,
0.5255, 0.5332,
        0.5285, 0.5296, 0.5293, 0.5294, 0.5294, 0.5294, 0.5294, 0.5255, 0.5332,
        0.5285, 0.5296, 0.5293, 0.5255, 0.5332, 0.5285, 0.5657])
finalReturns:  tensor([0.3784, 0.4145, 0.3853, 0.2996, 0.1803])
----------------------------------------
iter  2  stage  19  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([19, 19, 19, 25, 20, 19, 20, 19, 19, 19, 19, 19, 23, 19, 20,
19, 20, 20,
        19, 20, 20, 19, 19, 19,  0])
loss=  tensor(2.0426, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.6633,
```

```
2.6633, 2.6633, 2.6633, 2.6633, 2.6633, 2.6633, 2.6633, 2.6633,
        2.6633, 2.6633, 2.6633, 2.6633, 2.6633, 2.6633, 2.6633, 2.6633, 2.6633,
        2.6633, 2.6633, 2.0985, 1.6020, 1.1580, 0.7504, 0.3663]) return=
132091.35601931607
probs of actions:  tensor([0.7048, 0.6305, 0.5994, 0.0062, 0.2655, 0.6701,
0.2830, 0.6955, 0.7536,
        0.6773, 0.6796, 0.6751, 0.0043, 0.6330, 0.3343, 0.6932, 0.2750, 0.2573,
        0.5974, 0.3659, 0.3144, 0.5819, 0.7250, 0.3477, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4751, 0.5434, 0.5259, 0.5039, 0.5481, 0.5276, 0.5260,
0.5331, 0.5285,
        0.5296, 0.5293, 0.5294, 0.5126, 0.5445, 0.5218, 0.5341, 0.5243, 0.5296,
        0.5322, 0.5248, 0.5294, 0.5322, 0.5287, 0.5296, 0.5655])
finalReturns:  tensor([0.5469, 0.5869, 0.5539, 0.4658, 0.3446, 0.1992])
----------------------------------------
iter  2  stage  18  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([20, 20, 20, 20, 19, 20, 20, 20, 20, 20, 20, 20, 19, 19, 20,
20, 19, 19,
        20, 20, 20, 20, 20, 20,  0])
loss=  tensor(0.9034, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.0040,
3.0040, 3.0040, 3.0040, 3.0040, 3.0040, 3.0040, 3.0040, 3.0040,
        3.0040, 3.0040, 3.0040, 3.0040, 3.0040, 3.0040, 3.0040, 3.0040, 3.0040,
        3.0040, 2.4383, 1.9420, 1.4980, 1.0904, 0.7094, 0.3476]) return=
132119.19007904181
probs of actions:  tensor([0.6944, 0.7625, 0.7787, 0.6775, 0.2659, 0.7191,
0.7076, 0.7213, 0.6555,
        0.7170, 0.7109, 0.7317, 0.2381, 0.2190, 0.7402, 0.7420, 0.2519, 0.2741,
        0.8336, 0.8034, 0.7274, 0.7529, 0.7599, 0.9213, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4712, 0.5433, 0.5248, 0.5294, 0.5322, 0.5248, 0.5294,
0.5283, 0.5286,
        0.5285, 0.5285, 0.5285, 0.5324, 0.5287, 0.5257, 0.5292, 0.5322, 0.5287,
        0.5257, 0.5292, 0.5283, 0.5286, 0.5285, 0.5285, 0.5685])
finalReturns:  tensor([0.7334, 0.7734, 0.7404, 0.6561, 0.5351, 0.3877, 0.2209])
----------------------------------------
iter  2  stage  17  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([20, 20, 20, 20, 20, 21, 20, 19, 25, 19, 19, 19, 19, 21, 20,
20, 20, 19,
        20, 20, 19, 20, 19, 20,  0])
loss=  tensor(4.4889, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.3606,
3.3606, 3.3606, 3.3606, 3.3606, 3.3606, 3.3606, 3.3606, 3.3606,
        3.3606, 3.3606, 3.3606, 3.3606, 3.3606, 3.3606, 3.3606, 3.3606, 3.3606,
        2.7921, 2.2965, 1.8489, 1.4398, 1.0574, 0.6917, 0.3419]) return=
132054.13451253538
probs of actions:  tensor([0.6917, 0.7543, 0.7701, 0.6788, 0.7036, 0.0511,
0.7061, 0.2013, 0.0152,
        0.2061, 0.2140, 0.1935, 0.1996, 0.0562, 0.7250, 0.7468, 0.7122, 0.2247,
        0.7869, 0.8139, 0.1817, 0.7292, 0.1659, 0.9044, 1.0000],
```

```
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4712, 0.5433, 0.5248, 0.5294, 0.5283, 0.5245, 0.5323,
0.5315, 0.5025,
        0.5523, 0.5237, 0.5308, 0.5290, 0.5215, 0.5330, 0.5274, 0.5288, 0.5323,
        0.5248, 0.5295, 0.5322, 0.5248, 0.5333, 0.5245, 0.5695])
finalReturns:  tensor([0.9104, 0.9465, 0.9173, 0.8354, 0.7124, 0.5700, 0.4023,
0.2276])
----------------------------------------
iter  2  stage  16  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([19, 20, 19, 20, 20, 20, 20, 20, 20, 20, 20, 20, 23, 20, 25,
20, 20, 21,
        20, 20, 27, 21, 20, 20,  0])
loss=  tensor(10.3548, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([3.6175, 3.6175, 3.6175, 3.6175, 3.6175, 3.6175, 3.6175, 3.6175, 3.6175,
        3.6175, 3.6175, 3.6175, 3.6175, 3.6175, 3.6175, 3.6175, 3.0539,
        2.5571, 2.1132, 1.7088, 1.3293, 0.9688, 0.6424, 0.3189]) return=
131844.90001474044
probs of actions:  tensor([0.1382, 0.7796, 0.0968, 0.7182, 0.7469, 0.7440,
0.7399, 0.7577, 0.7211,
        0.7426, 0.7502, 0.7602, 0.0066, 0.7539, 0.0349, 0.7833, 0.7792, 0.0772,
        0.7695, 0.8051, 0.0124, 0.0893, 0.8103, 0.9007, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4751, 0.5395, 0.5297, 0.5254, 0.5293, 0.5283, 0.5286,
0.5285, 0.5285,
        0.5285, 0.5285, 0.5285, 0.5156, 0.5399, 0.5032, 0.5482, 0.5236, 0.5256,
        0.5320, 0.5277, 0.4958, 0.5511, 0.5257, 0.5292, 0.5683])
finalReturns:  tensor([1.1615, 1.2015, 1.1727, 1.0846, 0.9614, 0.8450, 0.6544,
0.4551, 0.2495])
----------------------------------------
iter  2  stage  15  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([20, 20, 20, 20, 27, 20, 20, 21, 19, 20, 20, 20, 21, 20, 19,
25, 20, 20,
        20, 20, 20, 20, 25, 20,  0])
loss=  tensor(7.6280, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.9346,
3.9346, 3.9346, 3.9346, 3.9346, 3.9346, 3.9346, 3.9346, 3.9346,
        3.9346, 3.9346, 3.9346, 3.9346, 3.9346, 3.9346, 3.9346, 3.3696, 2.8732,
        2.4457, 2.0460, 1.6717, 1.3147, 0.9705, 0.6359, 0.3083]) return=
131802.06019900876
probs of actions:  tensor([0.7385, 0.7852, 0.7893, 0.7300, 0.0182, 0.7516,
0.7516, 0.0807, 0.0850,
        0.7540, 0.7621, 0.7685, 0.0843, 0.7577, 0.0585, 0.0655, 0.7731, 0.7445,
        0.7731, 0.7853, 0.7374, 0.7530, 0.0523, 0.8993, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4712, 0.5433, 0.5248, 0.5294, 0.4954, 0.5553, 0.5219,
0.5261, 0.5358,
        0.5239, 0.5297, 0.5282, 0.5245, 0.5323, 0.5315, 0.5025, 0.5484, 0.5236,
        0.5298, 0.5282, 0.5286, 0.5285, 0.5060, 0.5475, 0.5638])
finalReturns:  tensor([1.3723, 1.4348, 1.3828, 1.2867, 1.1566, 1.0027, 0.8312,
```

```
0.6468, 0.4754,
        0.2555])
---------------------------------------
iter  2  stage  14  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([20, 25, 25, 20, 19, 25, 21, 27, 21, 20, 21, 25, 25, 20, 19,
19, 19, 20,
        25, 25, 23, 25, 25, 25,  0])
loss=  tensor(24.3833, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([4.2216, 4.2216, 4.2216, 4.2216, 4.2216, 4.2216, 4.2216, 4.2216, 4.2216,
        4.2216, 4.2216, 4.2216, 4.2216, 4.2216, 4.2216, 3.6567, 3.1602, 2.7129,
        2.3005, 1.9135, 1.5480, 1.2121, 0.8942, 0.5849, 0.2885]) return=
131224.90587325243
probs of actions:  tensor([0.3167, 0.4112, 0.4304, 0.3126, 0.0764, 0.4228,
0.0819, 0.0535, 0.0741,
        0.3408, 0.0841, 0.4277, 0.4054, 0.3495, 0.0591, 0.0609, 0.0489, 0.3048,
        0.4070, 0.4643, 0.0152, 0.4746, 0.4893, 0.3663, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4712, 0.5208, 0.5213, 0.5437, 0.5287, 0.5032, 0.5441,
0.4945, 0.5514,
        0.5256, 0.5251, 0.5096, 0.5241, 0.5430, 0.5288, 0.5295, 0.5294, 0.5255,
        0.5068, 0.5248, 0.5299, 0.5138, 0.5231, 0.5207, 0.5838])
finalReturns:  tensor([1.5945, 1.6306, 1.5975, 1.5155, 1.4024, 1.2826, 1.1233,
0.9293, 0.7334,
        0.5197, 0.2953])
---------------------------------------
iter  2  stage  13  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([25, 25, 25, 25, 25, 25, 20, 25, 20, 25, 25, 20, 25, 25, 25,
25, 25, 25,
        20, 25, 25, 25, 27, 21,  0])
loss=  tensor(11.5393, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([4.3408, 4.3408, 4.3408, 4.3408, 4.3408, 4.3408, 4.3408, 4.3408, 4.3408,
        4.3408, 4.3408, 4.3408, 4.3408, 4.3408, 3.7524, 3.2614, 2.8326, 2.4489,
        2.0972, 1.7685, 1.4565, 1.1431, 0.8458, 0.5561, 0.2730]) return=
130632.29857987192
probs of actions:  tensor([0.8027, 0.7853, 0.8031, 0.8005, 0.7618, 0.7843,
0.0751, 0.7990, 0.0598,
        0.7952, 0.7745, 0.0666, 0.7593, 0.8055, 0.8090, 0.8378, 0.8045, 0.7969,
        0.0747, 0.8099, 0.7747, 0.8152, 0.0467, 0.0235, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5401, 0.5165, 0.5224, 0.5209, 0.5213, 0.5437,
0.5023, 0.5485,
        0.5011, 0.5263, 0.5424, 0.5026, 0.5259, 0.5200, 0.5215, 0.5211, 0.5212,
        0.5437, 0.5023, 0.5260, 0.5200, 0.5111, 0.5472, 0.5667])
finalReturns:  tensor([1.9858, 2.0483, 2.0193, 1.9266, 1.7891, 1.6197, 1.4047,
1.2144, 1.0019,
        0.7792, 0.5578, 0.2936])
---------------------------------------
iter  2  stage  12  ep  99999   adversary:  AdversaryModes.imitation_128
```

```
  actions:  tensor([25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 27, 25,
27, 25, 27,
        25, 25, 25, 25, 25, 25,  0])
loss=  tensor(16.6462, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([4.5332, 4.5332, 4.5332, 4.5332, 4.5332, 4.5332, 4.5332, 4.5332, 4.5332,
        4.5332, 4.5332, 4.5332, 4.5332, 3.9495, 3.4574, 3.0288, 2.6513, 2.3024,
        1.9819, 1.6745, 1.3840, 1.0988, 0.8192, 0.5435, 0.2707]) return=
130222.86739483997
probs of actions:  tensor([0.8502, 0.8415, 0.8612, 0.8445, 0.8176, 0.8402,
0.8292, 0.8532, 0.8535,
        0.8500, 0.8340, 0.8461, 0.8389, 0.1052, 0.8543, 0.0961, 0.8421, 0.1346,
        0.8162, 0.8575, 0.8188, 0.8667, 0.9187, 0.9012, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5401, 0.5165, 0.5224, 0.5209, 0.5213, 0.5212,
0.5212, 0.5212,
        0.5212, 0.5212, 0.5212, 0.5212, 0.5108, 0.5289, 0.5089, 0.5293, 0.5088,
        0.5294, 0.5192, 0.5217, 0.5211, 0.5212, 0.5212, 0.5837])
finalReturns:  tensor([2.2921, 2.3546, 2.3359, 2.2356, 2.1043, 1.9237, 1.7355,
1.5135, 1.2849,
        1.0484, 0.8069, 0.5614, 0.3130])
----------------------------------------
iter  2  stage  11  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([25, 25, 25, 25, 27, 25, 27, 25, 27, 25, 25, 25, 25, 27, 27,
27, 25, 25,
        27, 25, 25, 27, 25, 27,  0])
loss=  tensor(23.1371, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([4.7819, 4.7819, 4.7819, 4.7819, 4.7819, 4.7819, 4.7819, 4.7819, 4.7819,
        4.7819, 4.7819, 4.7819, 4.1977, 3.7057, 3.2771, 2.8934, 2.5475, 2.2274,
        1.9262, 1.6334, 1.3481, 1.0734, 0.8000, 0.5293, 0.2653]) return=
129983.17438300277
probs of actions:  tensor([0.8312, 0.8252, 0.8517, 0.8265, 0.1764, 0.8258,
0.1594, 0.8394, 0.1417,
        0.8382, 0.8193, 0.8525, 0.8306, 0.1470, 0.1561, 0.1345, 0.8147, 0.7903,
        0.1904, 0.8504, 0.8082, 0.1412, 0.9162, 0.0991, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5401, 0.5165, 0.5224, 0.5105, 0.5289, 0.5089,
0.5293, 0.5088,
        0.5294, 0.5192, 0.5217, 0.5211, 0.5108, 0.5185, 0.5165, 0.5274, 0.5196,
        0.5112, 0.5288, 0.5193, 0.5113, 0.5287, 0.5089, 0.5918])
finalReturns:  tensor([2.5538, 2.6163, 2.5872, 2.5050, 2.3702, 2.1996, 1.9923,
1.7738, 1.5555,
        1.3120, 1.0674, 0.8294, 0.5714, 0.3265])
----------------------------------------
iter  2  stage  10  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([25, 27, 25, 25, 27, 25, 27, 27, 25, 25, 27, 27, 25, 25, 25,
25, 25, 27,
        27, 25, 25, 27, 25, 27,  0])
loss=  tensor(19.3476, grad_fn=<NegBackward0>)   ,  base rewards=
```

```
tensor([5.0490, 5.0490, 5.0490, 5.0490, 5.0490, 5.0490, 5.0490, 5.0490, 5.0490,
        5.0490, 5.0490, 4.4650, 3.9730, 3.5509, 3.1764, 2.8302, 2.5058, 2.1969,
        1.8996, 1.6107, 1.3334, 1.0633, 0.7937, 0.5257, 0.2638]) return=
129940.02348244
probs of actions:  tensor([0.6411, 0.3478, 0.6864, 0.6324, 0.3795, 0.6402,
0.3471, 0.3273, 0.6574,
        0.6662, 0.3993, 0.2574, 0.6499, 0.6282, 0.6171, 0.6509, 0.5852, 0.4329,
        0.4035, 0.6756, 0.6048, 0.3174, 0.7760, 0.2519, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5297, 0.5242, 0.5205, 0.5110, 0.5288, 0.5089,
0.5189, 0.5268,
        0.5198, 0.5111, 0.5184, 0.5270, 0.5198, 0.5216, 0.5211, 0.5212, 0.5108,
        0.5185, 0.5269, 0.5198, 0.5112, 0.5288, 0.5089, 0.5918])
finalReturns:  tensor([2.8077, 2.8806, 2.8543, 2.7493, 2.6041, 2.4288, 2.2321,
2.0197, 1.8063,
        1.5767, 1.3270, 1.0774, 0.8358, 0.5750, 0.3281])
----------------------------------------
iter  2  stage  9  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([27, 27, 27, 27, 27, 25, 25, 27, 25, 25, 27, 27, 25, 25, 25,
27, 25, 27,
        27, 25, 25, 27, 25, 27,  0])
loss=  tensor(23.0461, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([5.3010, 5.3010, 5.3010, 5.3010, 5.3010, 5.3010, 5.3010, 5.3010, 5.3010,
        5.3010, 4.7192, 4.2266, 3.7982, 3.4206, 3.0777, 2.7543, 2.4466, 2.1500,
        1.8670, 1.5875, 1.3174, 1.0525, 0.7869, 0.5219, 0.2621]) return=
129813.23890870859
probs of actions:  tensor([0.5159, 0.5150, 0.4584, 0.5240, 0.5492, 0.4800,
0.4801, 0.4968, 0.4940,
        0.4976, 0.5986, 0.4028, 0.4706, 0.4470, 0.4710, 0.5364, 0.4000, 0.5993,
        0.5524, 0.4923, 0.4229, 0.4827, 0.6139, 0.3995, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5375, 0.5118, 0.5182, 0.5166, 0.5274, 0.5196,
0.5112, 0.5288,
        0.5193, 0.5113, 0.5183, 0.5270, 0.5198, 0.5216, 0.5107, 0.5289, 0.5089,
        0.5189, 0.5268, 0.5198, 0.5111, 0.5288, 0.5089, 0.5918])
finalReturns:  tensor([3.0709, 3.1334, 3.1147, 3.0248, 2.8754, 2.6986, 2.5004,
2.2974, 2.0651,
        1.8392, 1.5998, 1.3431, 1.0881, 0.8426, 0.5789, 0.3297])
----------------------------------------
iter  2  stage  8  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([25, 27, 27, 27, 27, 27, 27, 25, 25, 27, 25, 25, 27, 25, 25,
27, 25, 27,
        27, 27, 27, 27, 27, 27,  0])
loss=  tensor(31.1249, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([5.5427, 5.5427, 5.5427, 5.5427, 5.5427, 5.5427, 5.5427, 5.5427, 5.5427,
        4.9605, 4.4680, 4.0396, 3.6620, 3.3132, 2.9870, 2.6824, 2.3868, 2.0995,
        1.8233, 1.5489, 1.2825, 1.0205, 0.7620, 0.5062, 0.2524]) return=
129679.56150960914
```

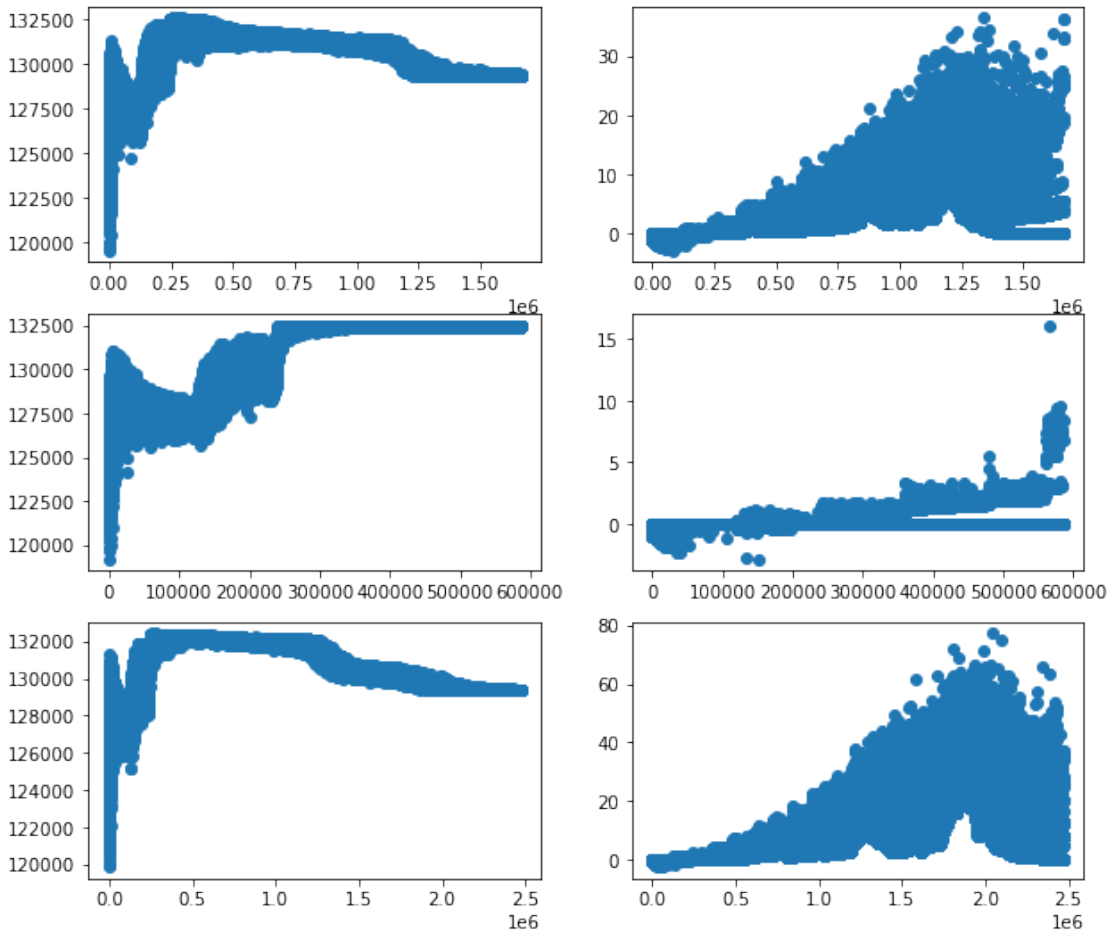
```
probs of actions:  tensor([0.2493, 0.7420, 0.6959, 0.7517, 0.7673, 0.7372,
0.7310, 0.2662, 0.2334,
        0.7474, 0.1887, 0.3415, 0.7633, 0.2293, 0.2571, 0.7751, 0.1964, 0.8103,
        0.7621, 0.7381, 0.8030, 0.7262, 0.6998, 0.6726, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4487, 0.5297, 0.5138, 0.5177, 0.5167, 0.5170, 0.5169,
0.5273, 0.5197,
        0.5112, 0.5288, 0.5193, 0.5113, 0.5287, 0.5193, 0.5113, 0.5287, 0.5089,
        0.5189, 0.5164, 0.5170, 0.5169, 0.5169, 0.5169, 0.5898])
finalReturns:  tensor([3.3375, 3.4000, 3.3813, 3.2810, 3.1392, 2.9768, 2.7742,
2.5595, 2.3439,
        2.1024, 1.8696, 1.6252, 1.3751, 1.1201, 0.8616, 0.6005, 0.3374])
----------------------------------------
iter  2  stage  7  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([27, 27, 27, 27, 25, 27, 27, 27, 27, 27, 27, 25, 27, 27, 27,
27, 27, 27,
        27, 27, 25, 27, 27, 27,  0])
loss=  tensor(15.3898, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([5.7261, 5.7261, 5.7261, 5.7261, 5.7261, 5.7261, 5.7261, 5.7261, 5.1368,
        4.6459, 4.2236, 3.8492, 3.5088, 3.1929, 2.8892, 2.6013, 2.3234, 2.0532,
        1.7886, 1.5282, 1.2709, 1.0159, 0.7627, 0.5058, 0.2524]) return=
129467.4131065994
probs of actions:  tensor([0.9444, 0.9402, 0.9242, 0.9438, 0.0534, 0.9387,
0.9331, 0.9445, 0.9529,
        0.9536, 0.9645, 0.0761, 0.9519, 0.9444, 0.9565, 0.9595, 0.9607, 0.9553,
        0.9569, 0.9400, 0.0355, 0.9364, 0.9489, 0.9312, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5375, 0.5118, 0.5182, 0.5270, 0.5093, 0.5188,
0.5164, 0.5170,
        0.5169, 0.5169, 0.5273, 0.5093, 0.5188, 0.5164, 0.5170, 0.5169, 0.5169,
        0.5169, 0.5169, 0.5273, 0.5093, 0.5188, 0.5164, 0.5899])
finalReturns:  tensor([3.6596, 3.7325, 3.7063, 3.6117, 3.4692, 3.2823, 3.0890,
2.8738, 2.6452,
        2.4061, 2.1594, 1.9071, 1.6506, 1.3909, 1.1186, 0.8625, 0.6006, 0.3376])
----------------------------------------
iter  2  stage  6  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,
27, 27, 27,
        27, 27, 27, 27, 27, 27,  0])
loss=  tensor(0.5009, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.9498,
5.9498, 5.9498, 5.9498, 5.9498, 5.9498, 5.9498, 5.3600, 4.8693,
        4.4470, 4.0725, 3.7322, 3.4162, 3.1181, 2.8329, 2.5572, 2.2886, 2.0252,
        1.7657, 1.5091, 1.2547, 1.0019, 0.7503, 0.4996, 0.2495]) return=
129338.69866666665
probs of actions:  tensor([0.9869, 0.9858, 0.9812, 0.9867, 0.9870, 0.9850,
0.9836, 0.9897, 0.9896,
        0.9908, 0.9930, 0.9842, 0.9890, 0.9900, 0.9919, 0.9932, 0.9919, 0.9909,
        0.9907, 0.9878, 0.9923, 0.9858, 0.9907, 0.9860, 1.0000],
```

```
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5375, 0.5118, 0.5182, 0.5166, 0.5170, 0.5169,
0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5898])
finalReturns:  tensor([3.9446, 4.0175, 3.9913, 3.8967, 3.7542, 3.5777, 3.3767,
3.1579, 2.9262,
        2.6849, 2.4366, 2.1831, 1.9257, 1.6653, 1.4028, 1.1387, 0.8734, 0.6072,
        0.3403])
----------------------------------------
iter  2  stage  5  ep  99999   adversary:  AdversaryModes.imitation_128
  actions:  tensor([27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,
27, 27, 27,
        27, 27, 27, 27, 27, 27,  0])
loss=  tensor(0.1744, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.1990,
6.1990, 6.1990, 6.1990, 6.1990, 6.1990, 5.6091, 5.1185, 4.6961,
        4.3217, 3.9813, 3.6654, 3.3672, 3.0820, 2.8063, 2.5377, 2.2744, 2.0149,
        1.7583, 1.5038, 1.2510, 0.9994, 0.7487, 0.4987, 0.2491]) return=
129338.69866666665
probs of actions:  tensor([0.9952, 0.9947, 0.9929, 0.9951, 0.9951, 0.9953,
0.9948, 0.9969, 0.9974,
        0.9973, 0.9979, 0.9956, 0.9958, 0.9967, 0.9979, 0.9983, 0.9980, 0.9967,
        0.9975, 0.9965, 0.9977, 0.9943, 0.9977, 0.9957, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5375, 0.5118, 0.5182, 0.5166, 0.5170, 0.5169,
0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5898])
finalReturns:  tensor([4.2124, 4.2853, 4.2591, 4.1645, 4.0220, 3.8455, 3.6445,
3.4257, 3.1940,
        2.9527, 2.7044, 2.4509, 2.1934, 1.9331, 1.6706, 1.4065, 1.1412, 0.8750,
        0.6081, 0.3407])
----------------------------------------
iter  2  stage  4  ep  66024   adversary:  AdversaryModes.imitation_128
  actions:  tensor([27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,
27, 27, 27,
        27, 27, 27, 27, 27, 27,  0])
loss=  tensor(0.0306, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.4476,
6.4476, 6.4476, 6.4476, 6.4476, 5.8581, 5.3673, 4.9450, 4.5705,
        4.2302, 3.9142, 3.6161, 3.3309, 3.0552, 2.7866, 2.5232, 2.2637, 2.0071,
        1.7527, 1.4999, 1.2483, 0.9976, 0.7475, 0.4980, 0.2489]) return=
129338.69866666665
probs of actions:  tensor([0.9990, 0.9988, 0.9984, 0.9989, 0.9990, 0.9993,
0.9992, 0.9996, 0.9996,
        0.9996, 0.9998, 0.9994, 0.9993, 0.9995, 0.9997, 0.9999, 0.9997, 0.9995,
        0.9997, 0.9994, 0.9998, 0.9989, 0.9998, 0.9994, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5375, 0.5118, 0.5182, 0.5166, 0.5170, 0.5169,
```

```
0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5898])
finalReturns:  tensor([4.4805, 4.5534, 4.5271, 4.4326, 4.2901, 4.1135, 3.9125,
3.6938, 3.4621,
        3.2208, 2.9725, 2.7189, 2.4615, 2.2012, 1.9387, 1.6746, 1.4093, 1.1430,
        0.8761, 0.6087, 0.3410])
---------------------------------------
iter  2  stage  3  ep  370   adversary:  AdversaryModes.imitation_128
  actions:  tensor([27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,
27, 27, 27,
        27, 27, 27, 27, 27, 27,  0])
loss=  tensor(0.0339, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.6975,
6.6975, 6.6975, 6.6975, 6.1064, 5.6160, 5.1936, 4.8192, 4.4788,
        4.1629, 3.8647, 3.5795, 3.3038, 3.0352, 2.7719, 2.5124, 2.2558, 2.0013,
        1.7485, 1.4969, 1.2462, 0.9962, 0.7466, 0.4975, 0.2486]) return=
129338.69866666665
probs of actions:  tensor([0.9990, 0.9989, 0.9985, 0.9990, 0.9991, 0.9993,
0.9993, 0.9996, 0.9996,
        0.9996, 0.9998, 0.9994, 0.9993, 0.9995, 0.9997, 0.9999, 0.9997, 0.9996,
        0.9998, 0.9994, 0.9998, 0.9990, 0.9998, 0.9994, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5375, 0.5118, 0.5182, 0.5166, 0.5170, 0.5169,
0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5898])
finalReturns:  tensor([4.7487, 4.8216, 4.7954, 4.7008, 4.5583, 4.3818, 4.1808,
3.9621, 3.7303,
        3.4891, 3.2408, 2.9872, 2.7298, 2.4694, 2.2070, 1.9429, 1.6775, 1.4113,
        1.1444, 0.8770, 0.6092, 0.3412])
---------------------------------------
iter  2  stage  2  ep  4830   adversary:  AdversaryModes.imitation_128
  actions:  tensor([27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,
27, 27, 27,
        27, 27, 27, 27, 27, 27,  0])
loss=  tensor(0.0356, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.9408,
6.9408, 6.9408, 6.3561, 5.8642, 5.4422, 5.0677, 4.7273, 4.4114,
        4.1132, 3.8280, 3.5523, 3.2837, 3.0203, 2.7608, 2.5042, 2.2498, 1.9970,
        1.7454, 1.4947, 1.2447, 0.9951, 0.7460, 0.4971, 0.2485]) return=
129338.69866666665
probs of actions:  tensor([0.9991, 0.9990, 0.9990, 0.9990, 0.9991, 0.9995,
0.9994, 0.9997, 0.9997,
        0.9997, 0.9999, 0.9995, 0.9994, 0.9996, 0.9998, 0.9999, 0.9997, 0.9996,
        0.9998, 0.9994, 0.9998, 0.9990, 0.9998, 0.9995, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5375, 0.5118, 0.5182, 0.5166, 0.5170, 0.5169,
0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169,
```

```
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5898])
finalReturns:  tensor([5.0173, 5.0902, 5.0638, 4.9693, 4.8268, 4.6503, 4.4493,
4.2305, 3.9988,
        3.7575, 3.5092, 3.2557, 2.9982, 2.7379, 2.4754, 2.2113, 1.9460, 1.6798,
        1.4129, 1.1455, 0.8777, 0.6096, 0.3413])
---------------------------------------
iter  2  stage  1  ep  10   adversary:  AdversaryModes.imitation_128
  actions:  tensor([27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,
27, 27, 27,
        27, 27, 27, 27, 27, 27,  0])
loss=  tensor(0.0417, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.2101,
7.2101, 6.5997, 6.1137, 5.6903, 5.3161, 4.9756, 4.6597, 4.3615,
        4.0763, 3.8007, 3.5321, 3.2687, 3.0092, 2.7526, 2.4982, 2.2454, 1.9938,
        1.7430, 1.4930, 1.2435, 0.9943, 0.7455, 0.4968, 0.2484]) return=
129338.69866666665
probs of actions:  tensor([0.9991, 0.9990, 0.9990, 0.9990, 0.9991, 0.9995,
0.9994, 0.9997, 0.9997,
        0.9997, 0.9999, 0.9995, 0.9994, 0.9996, 0.9998, 0.9999, 0.9997, 0.9996,
        0.9998, 0.9994, 0.9998, 0.9990, 0.9998, 0.9995, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5375, 0.5118, 0.5182, 0.5166, 0.5170, 0.5169,
0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5898])
finalReturns:  tensor([5.2854, 5.3583, 5.3326, 5.2378, 5.0954, 4.9188, 4.7178,
4.4991, 4.2673,
        4.0261, 3.7778, 3.5242, 3.2668, 3.0064, 2.7440, 2.4799, 2.2145, 1.9483,
        1.6814, 1.4140, 1.1462, 0.8782, 0.6099, 0.3415])
---------------------------------------
iter  2  stage  0  ep  0   adversary:  AdversaryModes.imitation_128
  actions:  tensor([27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,
27, 27, 27,
        27, 27, 27, 27, 27, 27,  0])
loss=  tensor(0.0479, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.3783,
6.8670, 6.3576, 5.9396, 5.5641, 5.2240, 4.9080, 4.6098, 4.3246,
        4.0490, 3.7804, 3.5170, 3.2575, 3.0009, 2.7465, 2.4936, 2.2420, 1.9913,
        1.7413, 1.4918, 1.2426, 0.9938, 0.7451, 0.4966, 0.2483]) return=
129338.69866666665
probs of actions:  tensor([0.9991, 0.9990, 0.9990, 0.9990, 0.9991, 0.9995,
0.9994, 0.9997, 0.9997,
        0.9997, 0.9999, 0.9995, 0.9994, 0.9996, 0.9998, 0.9999, 0.9997, 0.9996,
        0.9998, 0.9994, 0.9998, 0.9990, 0.9998, 0.9995, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4383, 0.5375, 0.5118, 0.5182, 0.5166, 0.5170, 0.5169,
0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169,
        0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5169, 0.5898])
finalReturns:  tensor([5.5556, 5.6285, 5.6005, 5.5067, 5.3639, 5.1875, 4.9865,
```

```
4.7677, 4.5360,
        4.2947, 4.0464, 3.7929, 3.5354, 3.2751, 3.0126, 2.7485, 2.4832, 2.2170,
        1.9501, 1.6827, 1.4149, 1.1468, 0.8786, 0.6101, 0.3416])
0,[1e-05,1][1, 10000, 1, 1],1682912335 saved
[2471239, 'tensor([0., 0., 0., 0., 1., 0., 0., 0., 0., 0., 0., 0.])',
129338.69866666665, 75332.17066666666, 0.047861263155937195, 1e-05, 1, 0,
'tensor([27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,
27,\n        27, 27, 27, 27, 27, 27,  0])', '[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.\n 1.]', '0,[1e-05,1][1, 10000, 1,
1],1682912335', 25, 50, 172677.5835126837, 219861.0138117109, 91586.68399118823,
132334.4363268229, 129338.69866666666, 72150.91909080048, 72150.91909080048,
88030.91902269641, 88030.91902269641, 105848.51071772553, 72150.91909080048,
88030.91902269641]
```




```
policy reset
----------------------------------------
iter  0  stage  24  ep  99999   adversary:  AdversaryModes.fight_132
  actions:  tensor([ 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
```

```
0,  0,  0,
         0,  0,  0,  0,  0, 11,  0])
loss=  tensor(-0., grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.6045,
0.6045, 0.6045, 0.6045, 0.6045, 0.6045, 0.6045, 0.6045, 0.6045,
        0.6045, 0.6045, 0.6045, 0.6045, 0.6045, 0.6045, 0.6045, 0.6045, 0.6045,
        0.6045, 0.6045, 0.6045, 0.6045, 0.6045, 0.6045, 0.6045]) return=
138853.62351012303
probs of actions:  tensor([0.8900, 0.8782, 0.9039, 0.8629, 0.8870, 0.8853,
0.8782, 0.8892, 0.8739,
        0.9083, 0.9194, 0.8914, 0.9129, 0.9059, 0.8999, 0.8765, 0.8980, 0.8966,
        0.8917, 0.8911, 0.8851, 0.8940, 0.9132, 0.0012, 0.9840],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5112, 0.5238, 0.5334, 0.5406, 0.5460, 0.5501, 0.5532,
0.5555, 0.5573,
        0.5586, 0.5595, 0.5603, 0.5608, 0.5613, 0.5616, 0.5618, 0.5620, 0.5621,
        0.5622, 0.5623, 0.5623, 0.5624, 0.5624, 0.5503, 0.6045])
finalReturns:  tensor([0.])
----------------------------------------
iter  0  stage  23  ep  99999   adversary:  AdversaryModes.fight_132
  actions:  tensor([8, 9, 8, 8, 9, 9, 8, 1, 8, 8, 8, 9, 8, 9, 8, 0, 0, 0, 5, 0,
8, 9, 8, 5,
        0])
loss=  tensor(0.0605, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.8075,
0.8075, 0.8075, 0.8075, 0.8075, 0.8075, 0.8075, 0.8075, 0.8075,
        0.8075, 0.8075, 0.8075, 0.8075, 0.8075, 0.8075, 0.8075, 0.8075, 0.8075,
        0.8075, 0.8075, 0.8075, 0.8075, 0.8075, 0.8075, 0.3942]) return=
111874.68821578428
probs of actions:  tensor([0.3709, 0.2137, 0.2994, 0.3707, 0.2204, 0.1786,
0.3582, 0.0190, 0.2927,
        0.3360, 0.3653, 0.2350, 0.3832, 0.1780, 0.3652, 0.2011, 0.1929, 0.2182,
        0.0257, 0.2460, 0.3083, 0.1934, 0.2999, 0.0109, 0.9932],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5048, 0.5451, 0.5831, 0.5434, 0.5128, 0.4952, 0.4839,
0.4771, 0.4374,
        0.4362, 0.4353, 0.4329, 0.4374, 0.4345, 0.4386, 0.4435, 0.4161, 0.3962,
        0.3790, 0.3861, 0.3677, 0.3817, 0.3986, 0.4109, 0.4100])
finalReturns:  tensor([0.0134, 0.0159])
----------------------------------------
iter  0  stage  22  ep  99999   adversary:  AdversaryModes.fight_132
  actions:  tensor([ 9, 11, 15, 15, 15, 15, 13, 11, 11, 11,  8, 11,  9, 16, 10,
11, 11, 14,
        11,  8,  0,  9, 15, 15,  0])
loss=  tensor(0.2034, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.2501,
1.2501, 1.2501, 1.2501, 1.2501, 1.2501, 1.2501, 1.2501, 1.2501,
        1.2501, 1.2501, 1.2501, 1.2501, 1.2501, 1.2501, 1.2501, 1.2501, 1.2501,
        1.2501, 1.2501, 1.2501, 1.2501, 1.2501, 0.8086, 0.3944]) return=
123413.1590088095
probs of actions:  tensor([0.1262, 0.2931, 0.2773, 0.2741, 0.2788, 0.2340,
```

```
0.0820, 0.3281, 0.2858,
        0.3215, 0.0544, 0.3065, 0.1199, 0.0377, 0.0284, 0.3071, 0.3022, 0.0157,
        0.3336, 0.0777, 0.0206, 0.1411, 0.4354, 0.3270, 0.9987],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.5448, 0.5776, 0.5443, 0.5482, 0.5390, 0.5378,
0.5301, 0.5136,
        0.5014, 0.4980, 0.4750, 0.4766, 0.4504, 0.4838, 0.4741, 0.4720, 0.4629,
        0.4796, 0.4818, 0.4751, 0.4305, 0.4190, 0.4414, 0.4811])
finalReturns:  tensor([0.0914, 0.1139, 0.0867])
----------------------------------------
iter  0  stage  21  ep  99999   adversary:  AdversaryModes.fight_132
  actions:  tensor([15, 15, 15, 15,  8, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0145, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.5548,
1.5548, 1.5548, 1.5548, 1.5548, 1.5548, 1.5548, 1.5548, 1.5548, 1.5548,
        1.5548, 1.5548, 1.5548, 1.5548, 1.5548, 1.5548, 1.5548, 1.5548, 1.5548,
        1.5548, 1.5548, 1.5548, 1.5548, 1.1058, 0.7056, 0.3403]) return=
111734.54075265459
probs of actions:  tensor([0.9539, 0.9634, 0.9569, 0.9437, 0.0011, 0.9408,
0.9508, 0.9561, 0.9347,
        0.9630, 0.9612, 0.9563, 0.9605, 0.9461, 0.9509, 0.9305, 0.9621, 0.9472,
        0.9524, 0.9441, 0.9524, 0.9738, 0.9814, 0.9642, 0.9999],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5570, 0.5228, 0.4978, 0.4956, 0.4418, 0.4379,
0.4350, 0.4329,
        0.4312, 0.4300, 0.4291, 0.4284, 0.4279, 0.4275, 0.4273, 0.4270, 0.4269,
        0.4268, 0.4267, 0.4266, 0.4266, 0.4265, 0.4265, 0.4490])
finalReturns:  tensor([0.1737, 0.1962, 0.1698, 0.1087])
----------------------------------------
iter  0  stage  20  ep  99999   adversary:  AdversaryModes.fight_132
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0103, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.8779,
1.8779, 1.8779, 1.8779, 1.8779, 1.8779, 1.8779, 1.8779, 1.8779, 1.8779,
        1.8779, 1.8779, 1.8779, 1.8779, 1.8779, 1.8779, 1.8779, 1.8779, 1.8779,
        1.8779, 1.8779, 1.8779, 1.4285, 1.0281, 0.6626, 0.3222]) return=
112524.66365021843
probs of actions:  tensor([0.9843, 0.9883, 0.9852, 0.9815, 0.9834, 0.9801,
0.9822, 0.9850, 0.9790,
        0.9890, 0.9879, 0.9852, 0.9864, 0.9812, 0.9827, 0.9752, 0.9875, 0.9837,
        0.9836, 0.9815, 0.9887, 0.9899, 0.9930, 0.9900, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5570, 0.5228, 0.4978, 0.4795, 0.4659, 0.4559,
0.4484, 0.4429,
        0.4387, 0.4356, 0.4333, 0.4316, 0.4303, 0.4293, 0.4286, 0.4280, 0.4276,
        0.4273, 0.4271, 0.4269, 0.4268, 0.4267, 0.4266, 0.4491])
```

```
finalReturns:  tensor([0.2781, 0.3006, 0.2743, 0.2131, 0.1269])
----------------------------------------
iter  0  stage  19  ep  82452   adversary:  AdversaryModes.fight_132
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0011, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.1872,
2.1872, 2.1872, 2.1872, 2.1872, 2.1872, 2.1872, 2.1872, 2.1872, 2.1872,
        2.1872, 2.1872, 2.1872, 2.1872, 2.1872, 2.1872, 2.1872, 2.1872, 2.1872,
        2.1872, 2.1872, 1.7377, 1.3371, 0.9715, 0.6310, 0.3088]) return=
112524.66365021843
probs of actions:  tensor([0.9987, 0.9991, 0.9988, 0.9984, 0.9986, 0.9982,
0.9984, 0.9987, 0.9981,
        0.9991, 0.9991, 0.9987, 0.9989, 0.9983, 0.9985, 0.9978, 0.9990, 0.9986,
        0.9985, 0.9990, 0.9993, 0.9994, 0.9998, 0.9996, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5570, 0.5228, 0.4978, 0.4795, 0.4659, 0.4559,
0.4484, 0.4429,
        0.4387, 0.4356, 0.4333, 0.4316, 0.4303, 0.4293, 0.4286, 0.4280, 0.4276,
        0.4273, 0.4271, 0.4269, 0.4268, 0.4267, 0.4266, 0.4491])
finalReturns:  tensor([0.3959, 0.4184, 0.3921, 0.3309, 0.2447, 0.1402])
----------------------------------------
iter  0  stage  18  ep  358   adversary:  AdversaryModes.fight_132
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0017, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.4870,
2.4870, 2.4870, 2.4870, 2.4870, 2.4870, 2.4870, 2.4870, 2.4870,
        2.4870, 2.4870, 2.4870, 2.4870, 2.4870, 2.4870, 2.4870, 2.4870, 2.4870,
        2.4870, 2.0371, 1.6364, 1.2707, 0.9301, 0.6079, 0.2990]) return=
112524.66365021843
probs of actions:  tensor([0.9989, 0.9992, 0.9989, 0.9986, 0.9988, 0.9984,
0.9986, 0.9989, 0.9984,
        0.9992, 0.9992, 0.9989, 0.9990, 0.9985, 0.9986, 0.9981, 0.9991, 0.9988,
        0.9990, 0.9991, 0.9994, 0.9994, 0.9998, 0.9996, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5570, 0.5228, 0.4978, 0.4795, 0.4659, 0.4559,
0.4484, 0.4429,
        0.4387, 0.4356, 0.4333, 0.4316, 0.4303, 0.4293, 0.4286, 0.4280, 0.4276,
        0.4273, 0.4271, 0.4269, 0.4268, 0.4267, 0.4266, 0.4491])
finalReturns:  tensor([0.5235, 0.5460, 0.5197, 0.4585, 0.3722, 0.2678, 0.1501])
----------------------------------------
iter  0  stage  17  ep  94   adversary:  AdversaryModes.fight_132
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0026, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.7797,
2.7797, 2.7797, 2.7797, 2.7797, 2.7797, 2.7797, 2.7797, 2.7797,
```

```
        2.7797, 2.7797, 2.7797, 2.7797, 2.7797, 2.7797, 2.7797, 2.7797, 2.7797,
        2.3296, 1.9286, 1.5627, 1.2221, 0.8997, 0.5908, 0.2918]) return=
112524.66365021843
probs of actions:  tensor([0.9989, 0.9992, 0.9990, 0.9987, 0.9988, 0.9985,
0.9987, 0.9989, 0.9984,
        0.9993, 0.9992, 0.9990, 0.9991, 0.9986, 0.9987, 0.9982, 0.9992, 0.9990,
        0.9991, 0.9992, 0.9994, 0.9995, 0.9998, 0.9996, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5570, 0.5228, 0.4978, 0.4795, 0.4659, 0.4559,
0.4484, 0.4429,
        0.4387, 0.4356, 0.4333, 0.4316, 0.4303, 0.4293, 0.4286, 0.4280, 0.4276,
        0.4273, 0.4271, 0.4269, 0.4268, 0.4267, 0.4266, 0.4491])
finalReturns:  tensor([0.6584, 0.6809, 0.6545, 0.5933, 0.5071, 0.4026, 0.2849,
0.1573])
----------------------------------------
iter  0  stage  16  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0036, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.0674,
3.0674, 3.0674, 3.0674, 3.0674, 3.0674, 3.0674, 3.0674, 3.0674,
        3.0674, 3.0674, 3.0674, 3.0674, 3.0674, 3.0674, 3.0674, 3.0674, 2.6169,
        2.2157, 1.8496, 1.5087, 1.1863, 0.8773, 0.5782, 0.2864]) return=
112524.66365021843
probs of actions:  tensor([0.9989, 0.9992, 0.9990, 0.9987, 0.9988, 0.9985,
0.9987, 0.9989, 0.9984,
        0.9993, 0.9992, 0.9990, 0.9991, 0.9986, 0.9987, 0.9982, 0.9992, 0.9990,
        0.9991, 0.9992, 0.9994, 0.9995, 0.9998, 0.9996, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5570, 0.5228, 0.4978, 0.4795, 0.4659, 0.4559,
0.4484, 0.4429,
        0.4387, 0.4356, 0.4333, 0.4316, 0.4303, 0.4293, 0.4286, 0.4280, 0.4276,
        0.4273, 0.4271, 0.4269, 0.4268, 0.4267, 0.4266, 0.4491])
finalReturns:  tensor([0.7987, 0.8212, 0.7948, 0.7336, 0.6473, 0.5429, 0.4251,
0.2975, 0.1627])
----------------------------------------
iter  0  stage  15  ep  305   adversary:  AdversaryModes.fight_132
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0035, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.3516,
3.3516, 3.3516, 3.3516, 3.3516, 3.3516, 3.3516, 3.3516, 3.3516,
        3.3516, 3.3516, 3.3516, 3.3516, 3.3516, 3.3516, 3.3516, 2.9005, 2.4989,
        2.1326, 1.7916, 1.4690, 1.1598, 0.8606, 0.5688, 0.2824]) return=
112524.66365021843
probs of actions:  tensor([0.9992, 0.9994, 0.9993, 0.9990, 0.9991, 0.9989,
0.9990, 0.9992, 0.9988,
        0.9995, 0.9994, 0.9992, 0.9993, 0.9990, 0.9991, 0.9990, 0.9995, 0.9994,
```

```
        0.9994, 0.9994, 0.9996, 0.9996, 0.9999, 0.9997, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5570, 0.5228, 0.4978, 0.4795, 0.4659, 0.4559,
0.4484, 0.4429,
        0.4387, 0.4356, 0.4333, 0.4316, 0.4303, 0.4293, 0.4286, 0.4280, 0.4276,
        0.4273, 0.4271, 0.4269, 0.4268, 0.4267, 0.4266, 0.4491])
finalReturns:  tensor([0.9431, 0.9656, 0.9392, 0.8779, 0.7916, 0.6871, 0.5693,
0.4417, 0.3069,
        0.1667])
---------------------------------------
iter  0  stage  14  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0049, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.6335,
3.6335, 3.6335, 3.6335, 3.6335, 3.6335, 3.6335, 3.6335, 3.6335,
        3.6335, 3.6335, 3.6335, 3.6335, 3.6335, 3.6335, 3.1817, 2.7796, 2.4128,
        2.0716, 1.7488, 1.4395, 1.1402, 0.8483, 0.5618, 0.2794]) return=
112524.66365021843
probs of actions:  tensor([0.9992, 0.9994, 0.9993, 0.9990, 0.9992, 0.9989,
0.9990, 0.9992, 0.9988,
        0.9995, 0.9994, 0.9992, 0.9993, 0.9990, 0.9991, 0.9990, 0.9995, 0.9994,
        0.9994, 0.9994, 0.9996, 0.9996, 0.9999, 0.9997, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5570, 0.5228, 0.4978, 0.4795, 0.4659, 0.4559,
0.4484, 0.4429,
        0.4387, 0.4356, 0.4333, 0.4316, 0.4303, 0.4293, 0.4286, 0.4280, 0.4276,
        0.4273, 0.4271, 0.4269, 0.4268, 0.4267, 0.4266, 0.4491])
finalReturns:  tensor([1.0905, 1.1130, 1.0866, 1.0253, 0.9389, 0.8344, 0.7166,
0.5889, 0.4541,
        0.3139, 0.1697])
---------------------------------------
iter  0  stage  13  ep  9   adversary:  AdversaryModes.fight_132
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0065, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.9141,
3.9141, 3.9141, 3.9141, 3.9141, 3.9141, 3.9141, 3.9141, 3.9141,
        3.9141, 3.9141, 3.9141, 3.9141, 3.9141, 3.4613, 3.0585, 2.6912, 2.3496,
        2.0265, 1.7171, 1.4177, 1.1256, 0.8391, 0.5566, 0.2772]) return=
112524.66365021843
probs of actions:  tensor([0.9992, 0.9994, 0.9993, 0.9991, 0.9992, 0.9989,
0.9990, 0.9992, 0.9988,
        0.9995, 0.9995, 0.9992, 0.9993, 0.9990, 0.9991, 0.9990, 0.9995, 0.9994,
        0.9994, 0.9994, 0.9996, 0.9996, 0.9999, 0.9997, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5570, 0.5228, 0.4978, 0.4795, 0.4659, 0.4559,
0.4484, 0.4429,
```

```
        0.4387, 0.4356, 0.4333, 0.4316, 0.4303, 0.4293, 0.4286, 0.4280, 0.4276,
        0.4273, 0.4271, 0.4269, 0.4268, 0.4267, 0.4266, 0.4491])
finalReturns:  tensor([1.2403, 1.2628, 1.2363, 1.1749, 1.0885, 0.9839, 0.8661,
0.7384, 0.6036,
        0.4633, 0.3191, 0.1719])
----------------------------------------
iter  0  stage  12  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0079, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.1941,
4.1941, 4.1941, 4.1941, 4.1941, 4.1941, 4.1941, 4.1941, 4.1941,
        4.1941, 4.1941, 4.1941, 4.1941, 3.7400, 3.3363, 2.9684, 2.6263, 2.3029,
        1.9932, 1.6936, 1.4014, 1.1147, 0.8322, 0.5527, 0.2755]) return=
112524.66365021843
probs of actions:  tensor([0.9992, 0.9994, 0.9993, 0.9991, 0.9992, 0.9989,
0.9990, 0.9992, 0.9988,
        0.9995, 0.9995, 0.9992, 0.9993, 0.9990, 0.9991, 0.9990, 0.9995, 0.9994,
        0.9994, 0.9994, 0.9996, 0.9996, 0.9999, 0.9997, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5570, 0.5228, 0.4978, 0.4795, 0.4659, 0.4559,
0.4484, 0.4429,
        0.4387, 0.4356, 0.4333, 0.4316, 0.4303, 0.4293, 0.4286, 0.4280, 0.4276,
        0.4273, 0.4271, 0.4269, 0.4268, 0.4267, 0.4266, 0.4491])
finalReturns:  tensor([1.3918, 1.4143, 1.3877, 1.3263, 1.2399, 1.1352, 1.0173,
0.8896, 0.7547,
        0.6144, 0.4702, 0.3230, 0.1736])
----------------------------------------
iter  0  stage  11  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0095, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.4745,
4.4745, 4.4745, 4.4745, 4.4745, 4.4745, 4.4745, 4.4745, 4.4745,
        4.4745, 4.4745, 4.4745, 4.0187, 3.6137, 3.2449, 2.9022, 2.5783, 2.2683,
        1.9684, 1.6760, 1.3893, 1.1066, 0.8270, 0.5498, 0.2742]) return=
112524.66365021843
probs of actions:  tensor([0.9992, 0.9994, 0.9993, 0.9991, 0.9992, 0.9989,
0.9990, 0.9992, 0.9988,
        0.9995, 0.9995, 0.9992, 0.9993, 0.9990, 0.9991, 0.9990, 0.9996, 0.9994,
        0.9994, 0.9994, 0.9996, 0.9996, 0.9999, 0.9997, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5570, 0.5228, 0.4978, 0.4795, 0.4659, 0.4559,
0.4484, 0.4429,
        0.4387, 0.4356, 0.4333, 0.4316, 0.4303, 0.4293, 0.4286, 0.4280, 0.4276,
        0.4273, 0.4271, 0.4269, 0.4268, 0.4267, 0.4266, 0.4491])
finalReturns:  tensor([1.5447, 1.5672, 1.5406, 1.4791, 1.3925, 1.2878, 1.1698,
1.0421, 0.9071,
```

```
        0.7668, 0.6226, 0.4753, 0.3259, 0.1748])
----------------------------------------
iter  0  stage  10  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0108, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.7560,
4.7560, 4.7560, 4.7560, 4.7560, 4.7560, 4.7560, 4.7560, 4.7560,
        4.7560, 4.7560, 4.2979, 3.8913, 3.5214, 3.1778, 2.8533, 2.5428, 2.2426,
        1.9500, 1.6630, 1.3802, 1.1005, 0.8232, 0.5476, 0.2733]) return=
112524.66365021843
probs of actions:  tensor([0.9992, 0.9994, 0.9993, 0.9991, 0.9992, 0.9989,
0.9990, 0.9992, 0.9988,
        0.9995, 0.9995, 0.9992, 0.9993, 0.9990, 0.9991, 0.9990, 0.9996, 0.9994,
        0.9994, 0.9994, 0.9996, 0.9996, 0.9999, 0.9997, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5570, 0.5228, 0.4978, 0.4795, 0.4659, 0.4559,
0.4484, 0.4429,
        0.4387, 0.4356, 0.4333, 0.4316, 0.4303, 0.4293, 0.4286, 0.4280, 0.4276,
        0.4273, 0.4271, 0.4269, 0.4268, 0.4267, 0.4266, 0.4491])
finalReturns:  tensor([1.6988, 1.7213, 1.6946, 1.6330, 1.5463, 1.4415, 1.3234,
1.1955, 1.0605,
        0.9202, 0.7759, 0.6286, 0.4792, 0.3281, 0.1757])
----------------------------------------
iter  0  stage  9  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0122, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.0396,
5.0396, 5.0396, 5.0396, 5.0396, 5.0396, 5.0396, 5.0396, 5.0396,
        5.0396, 4.5784, 4.1697, 3.7981, 3.4534, 3.1281, 2.8170, 2.5164, 2.2234,
        1.9362, 1.6532, 1.3734, 1.0960, 0.8203, 0.5460, 0.2726]) return=
112524.66365021843
probs of actions:  tensor([0.9992, 0.9994, 0.9993, 0.9991, 0.9992, 0.9989,
0.9990, 0.9992, 0.9989,
        0.9995, 0.9995, 0.9993, 0.9993, 0.9990, 0.9991, 0.9990, 0.9996, 0.9994,
        0.9994, 0.9994, 0.9996, 0.9996, 0.9999, 0.9997, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5570, 0.5228, 0.4978, 0.4795, 0.4659, 0.4559,
0.4484, 0.4429,
        0.4387, 0.4356, 0.4333, 0.4316, 0.4303, 0.4293, 0.4286, 0.4280, 0.4276,
        0.4273, 0.4271, 0.4269, 0.4268, 0.4267, 0.4266, 0.4491])
finalReturns:  tensor([1.8539, 1.8764, 1.8496, 1.7878, 1.7009, 1.5959, 1.4777,
1.3498, 1.2147,
        1.0743, 0.9299, 0.7827, 0.6332, 0.4821, 0.3297, 0.1764])
----------------------------------------
iter  0  stage  8  ep  31   adversary:  AdversaryModes.fight_132
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
```

```
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0130, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.3265,
5.3265, 5.3265, 5.3265, 5.3265, 5.3265, 5.3265, 5.3265, 5.3265,
        4.8611, 4.4494, 4.0759, 3.7296, 3.4032, 3.0913, 2.7901, 2.4967, 2.2091,
        1.9259, 1.6459, 1.3684, 1.0926, 0.8182, 0.5448, 0.2721]) return=
112524.66365021843
probs of actions:  tensor([0.9993, 0.9995, 0.9994, 0.9991, 0.9992, 0.9990,
0.9991, 0.9993, 0.9990,
        0.9995, 0.9995, 0.9994, 0.9994, 0.9992, 0.9992, 0.9991, 0.9996, 0.9995,
        0.9995, 0.9995, 0.9996, 0.9997, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5570, 0.5228, 0.4978, 0.4795, 0.4659, 0.4559,
0.4484, 0.4429,
        0.4387, 0.4356, 0.4333, 0.4316, 0.4303, 0.4293, 0.4286, 0.4280, 0.4276,
        0.4273, 0.4271, 0.4269, 0.4268, 0.4267, 0.4266, 0.4491])
finalReturns:  tensor([2.0099, 2.0324, 2.0054, 1.9434, 1.8563, 1.7511, 1.6327,
1.5047, 1.3695,
        1.2290, 1.0846, 0.9372, 0.7877, 0.6366, 0.4842, 0.3309, 0.1770])
---------------------------------------
iter  0  stage  7  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0149, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.6180,
5.6180, 5.6180, 5.6180, 5.6180, 5.6180, 5.6180, 5.6180, 5.1471,
        4.7315, 4.3551, 4.0068, 3.6790, 3.3660, 3.0639, 2.7700, 2.4820, 2.1984,
        1.9182, 1.6405, 1.3646, 1.0900, 0.8166, 0.5438, 0.2717]) return=
112524.66365021843
probs of actions:  tensor([0.9993, 0.9995, 0.9994, 0.9991, 0.9992, 0.9990,
0.9991, 0.9993, 0.9990,
        0.9995, 0.9995, 0.9994, 0.9994, 0.9992, 0.9992, 0.9991, 0.9996, 0.9995,
        0.9995, 0.9995, 0.9996, 0.9997, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5570, 0.5228, 0.4978, 0.4795, 0.4659, 0.4559,
0.4484, 0.4429,
        0.4387, 0.4356, 0.4333, 0.4316, 0.4303, 0.4293, 0.4286, 0.4280, 0.4276,
        0.4273, 0.4271, 0.4269, 0.4268, 0.4267, 0.4266, 0.4491])
finalReturns:  tensor([2.1668, 2.1893, 2.1621, 2.0997, 2.0124, 1.9069, 1.7883,
1.6601, 1.5248,
        1.3842, 1.2397, 1.0923, 0.9427, 0.7915, 0.6391, 0.4858, 0.3318, 0.1774])
---------------------------------------
iter  0  stage  6  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0173, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.9160,
5.9160, 5.9160, 5.9160, 5.9160, 5.9160, 5.9160, 5.4376, 5.0168,
```

```
        4.6366, 4.2857, 3.9558, 3.6414, 3.3383, 3.0435, 2.7550, 2.4710, 2.1905,
        1.9125, 1.6364, 1.3617, 1.0881, 0.8153, 0.5432, 0.2714]) return=
112524.66365021843
probs of actions:  tensor([0.9993, 0.9995, 0.9994, 0.9991, 0.9992, 0.9990,
0.9991, 0.9993, 0.9990,
        0.9996, 0.9995, 0.9994, 0.9994, 0.9992, 0.9992, 0.9991, 0.9996, 0.9995,
        0.9995, 0.9995, 0.9996, 0.9997, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5570, 0.5228, 0.4978, 0.4795, 0.4659, 0.4559,
0.4484, 0.4429,
        0.4387, 0.4356, 0.4333, 0.4316, 0.4303, 0.4293, 0.4286, 0.4280, 0.4276,
        0.4273, 0.4271, 0.4269, 0.4268, 0.4267, 0.4266, 0.4491])
finalReturns:  tensor([2.3247, 2.3472, 2.3197, 2.2569, 2.1692, 2.0634, 1.9445,
1.8160, 1.6805,
        1.5398, 1.3951, 1.2477, 1.0980, 0.9468, 0.7944, 0.6410, 0.4870, 0.3325,
        0.1777])
----------------------------------------
iter  0  stage  5  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0201, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.2229,
6.2229, 6.2229, 6.2229, 6.2229, 6.2229, 5.7345, 5.3065, 4.9214,
        4.5668, 4.2343, 3.9180, 3.6135, 3.3177, 3.0283, 2.7438, 2.4628, 2.1845,
        1.9082, 1.6333, 1.3596, 1.0867, 0.8144, 0.5426, 0.2712]) return=
112524.66365021843
probs of actions:  tensor([0.9993, 0.9995, 0.9994, 0.9991, 0.9993, 0.9990,
0.9991, 0.9993, 0.9990,
        0.9996, 0.9995, 0.9994, 0.9994, 0.9992, 0.9992, 0.9991, 0.9996, 0.9995,
        0.9995, 0.9995, 0.9996, 0.9997, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5570, 0.5228, 0.4978, 0.4795, 0.4659, 0.4559,
0.4484, 0.4429,
        0.4387, 0.4356, 0.4333, 0.4316, 0.4303, 0.4293, 0.4286, 0.4280, 0.4276,
        0.4273, 0.4271, 0.4269, 0.4268, 0.4267, 0.4266, 0.4491])
finalReturns:  tensor([2.4838, 2.5063, 2.4783, 2.4151, 2.3268, 2.2205, 2.1012,
1.9724, 1.8367,
        1.6957, 1.5510, 1.4034, 1.2536, 1.1023, 0.9498, 0.7965, 0.6425, 0.4879,
        0.3330, 0.1779])
----------------------------------------
iter  0  stage  4  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0225, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.5419,
6.5419, 6.5419, 6.5419, 6.5419, 6.0399, 5.6025, 5.2106, 4.8512,
        4.5152, 4.1963, 3.8899, 3.5927, 3.3023, 3.0169, 2.7354, 2.4567, 2.1800,
        1.9049, 1.6310, 1.3580, 1.0856, 0.8138, 0.5423, 0.2710]) return=
```

```
112524.66365021843
probs of actions:  tensor([0.9993, 0.9995, 0.9994, 0.9992, 0.9993, 0.9990,
0.9991, 0.9993, 0.9990,
        0.9996, 0.9995, 0.9994, 0.9994, 0.9992, 0.9992, 0.9992, 0.9996, 0.9995,
        0.9995, 0.9995, 0.9996, 0.9997, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5570, 0.5228, 0.4978, 0.4795, 0.4659, 0.4559,
0.4484, 0.4429,
        0.4387, 0.4356, 0.4333, 0.4316, 0.4303, 0.4293, 0.4286, 0.4280, 0.4276,
        0.4273, 0.4271, 0.4269, 0.4268, 0.4267, 0.4266, 0.4491])
finalReturns:  tensor([2.6442, 2.6667, 2.6382, 2.5743, 2.4853, 2.3784, 2.2586,
2.1294, 1.9932,
        1.8521, 1.7071, 1.5593, 1.4095, 1.2581, 1.1056, 0.9521, 0.7981, 0.6435,
        0.4886, 0.3334, 0.1780])
----------------------------------------
iter  0  stage  3  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0253, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.8775,
6.8775, 6.8775, 6.8775, 6.3572, 5.9069, 5.5059, 5.1400, 4.7993,
        4.4769, 4.1680, 3.8689, 3.5771, 3.2908, 3.0084, 2.7291, 2.4521, 2.1767,
        1.9025, 1.6293, 1.3568, 1.0848, 0.8132, 0.5420, 0.2709]) return=
112524.66365021843
probs of actions:  tensor([0.9993, 0.9995, 0.9994, 0.9992, 0.9993, 0.9990,
0.9991, 0.9993, 0.9990,
        0.9996, 0.9995, 0.9994, 0.9995, 0.9992, 0.9992, 0.9992, 0.9996, 0.9995,
        0.9995, 0.9995, 0.9996, 0.9997, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5570, 0.5228, 0.4978, 0.4795, 0.4659, 0.4559,
0.4484, 0.4429,
        0.4387, 0.4356, 0.4333, 0.4316, 0.4303, 0.4293, 0.4286, 0.4280, 0.4276,
        0.4273, 0.4271, 0.4269, 0.4268, 0.4267, 0.4266, 0.4491])
finalReturns:  tensor([2.8065, 2.8290, 2.7997, 2.7348, 2.6449, 2.5372, 2.4167,
2.2869, 2.1503,
        2.0088, 1.8636, 1.7156, 1.5656, 1.4141, 1.2615, 1.1080, 0.9539, 0.7993,
        0.6443, 0.4891, 0.3337, 0.1782])
----------------------------------------
iter  0  stage  2  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0276, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.2357,
7.2357, 7.2357, 6.6905, 6.2228, 5.8096, 5.4348, 5.0878, 4.7608,
        4.4484, 4.1469, 3.8532, 3.5655, 3.2822, 3.0021, 2.7244, 2.4486, 2.1741,
        1.9007, 1.6280, 1.3559, 1.0842, 0.8129, 0.5417, 0.2708]) return=
112524.66365021843
probs of actions:  tensor([0.9993, 0.9995, 0.9994, 0.9992, 0.9993, 0.9990,
```

```
0.9991, 0.9993, 0.9990,
        0.9996, 0.9995, 0.9994, 0.9995, 0.9992, 0.9992, 0.9992, 0.9996, 0.9995,
        0.9995, 0.9995, 0.9996, 0.9997, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5570, 0.5228, 0.4978, 0.4795, 0.4659, 0.4559,
0.4484, 0.4429,
        0.4387, 0.4356, 0.4333, 0.4316, 0.4303, 0.4293, 0.4286, 0.4280, 0.4276,
        0.4273, 0.4271, 0.4269, 0.4268, 0.4267, 0.4266, 0.4491])
finalReturns:  tensor([2.9710, 2.9935, 2.9633, 2.8971, 2.8059, 2.6971, 2.5757,
2.4451, 2.3080,
        2.1660, 2.0204, 1.8722, 1.7220, 1.5703, 1.4175, 1.2640, 1.1098, 0.9551,
        0.8002, 0.6449, 0.4895, 0.3339, 0.1783])
----------------------------------------
iter  0  stage  1  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0295, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.6252,
7.6252, 7.0457, 6.5544, 6.1245, 5.7379, 5.3822, 5.0490, 4.7321,
        4.4272, 4.1311, 3.8415, 3.5568, 3.2757, 2.9973, 2.7209, 2.4460, 2.1723,
        1.8993, 1.6271, 1.3552, 1.0838, 0.8126, 0.5416, 0.2707]) return=
112524.66365021843
probs of actions:  tensor([0.9993, 0.9995, 0.9994, 0.9992, 0.9993, 0.9990,
0.9991, 0.9993, 0.9990,
        0.9996, 0.9996, 0.9994, 0.9995, 0.9992, 0.9993, 0.9992, 0.9996, 0.9995,
        0.9995, 0.9995, 0.9996, 0.9997, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5570, 0.5228, 0.4978, 0.4795, 0.4659, 0.4559,
0.4484, 0.4429,
        0.4387, 0.4356, 0.4333, 0.4316, 0.4303, 0.4293, 0.4286, 0.4280, 0.4276,
        0.4273, 0.4271, 0.4269, 0.4268, 0.4267, 0.4266, 0.4491])
finalReturns:  tensor([3.1385, 3.1610, 3.1295, 3.0617, 2.9688, 2.8585, 2.7358,
2.6043, 2.4664,
        2.3237, 2.1777, 2.0291, 1.8786, 1.7267, 1.5738, 1.4201, 1.2658, 1.1111,
        0.9561, 0.8008, 0.6454, 0.4898, 0.3341, 0.1783])
----------------------------------------
iter  0  stage  0  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0320, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.9425,
7.4312, 6.9074, 6.4547, 6.0520, 5.6848, 5.3432, 5.0202, 4.7108,
        4.4114, 4.1193, 3.8328, 3.5503, 3.2709, 2.9937, 2.7183, 2.4441, 2.1709,
        1.8983, 1.6263, 1.3547, 1.0834, 0.8124, 0.5415, 0.2707]) return=
112524.66365021843
probs of actions:  tensor([0.9993, 0.9995, 0.9994, 0.9992, 0.9993, 0.9991,
0.9992, 0.9993, 0.9990,
        0.9996, 0.9996, 0.9994, 0.9995, 0.9992, 0.9993, 0.9992, 0.9996, 0.9995,
```

```
        0.9995, 0.9995, 0.9996, 0.9997, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5570, 0.5228, 0.4978, 0.4795, 0.4659, 0.4559,
0.4484, 0.4429,
        0.4387, 0.4356, 0.4333, 0.4316, 0.4303, 0.4293, 0.4286, 0.4280, 0.4276,
        0.4273, 0.4271, 0.4269, 0.4268, 0.4267, 0.4266, 0.4491])
finalReturns:  tensor([3.3100, 3.3325, 3.2993, 3.2292, 3.1341, 3.0218, 2.8975,
2.7647, 2.6257,
        2.4822, 2.3355, 2.1864, 2.0356, 1.8834, 1.7303, 1.5764, 1.4221, 1.2673,
        1.1122, 0.9568, 0.8013, 0.6457, 0.4900, 0.3342, 0.1784])
0,[1e-05,1][1, 10000, 1, 1],1682929682 saved
[683269, 'tensor([0., 0., 0., 0., 0., 1., 0., 0., 0., 0., 0., 0.])',
112524.66365021843, 93537.02666495038, 0.03203679621219635, 1e-05, 1, 0,
'tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15,\n        15, 15, 15, 15, 15, 15,  0])', '[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.\n 1.]', '0,[1e-05,1][1, 10000, 1,
1],1682929682', 25, 50, 161287.58344151574, 184572.3091501231, 73453.3647159636,
135545.41866666666, 132647.75466666667, 112524.66365021843, 112524.66365021843,
117051.59338141435, 117058.92268888396, 85375.85726043607, 112524.66365021843,
117061.61105796597]
policy reset
----------------------------------------
iter  1  stage  24  ep  99999   adversary:  AdversaryModes.fight_132
  actions:  tensor([0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
0, 0, 0, 0,
        0])
loss=  tensor(-0., grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.4723,
0.4723, 0.4723, 0.4723, 0.4723, 0.4723, 0.4723, 0.4723, 0.4723,
        0.4723, 0.4723, 0.4723, 0.4723, 0.4723, 0.4723, 0.4723, 0.4723, 0.4723,
        0.4723, 0.4723, 0.4723, 0.4723, 0.4723, 0.4723, 0.4723]) return=
133459.1809763887
probs of actions:  tensor([0.9112, 0.0541, 0.9293, 0.9027, 0.9259, 0.9068,
0.9231, 0.9408, 0.9248,
        0.9224, 0.9068, 0.9271, 0.9228, 0.9337, 0.9150, 0.0406, 0.9252, 0.9250,
        0.8971, 0.9140, 0.9118, 0.9137, 0.9213, 0.9058, 0.9889],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5112, 0.5237, 0.5370, 0.5433, 0.5481, 0.5517, 0.5544,
0.5564, 0.5579,
        0.5591, 0.5599, 0.5606, 0.5610, 0.5614, 0.5617, 0.5618, 0.5658, 0.5390,
        0.5193, 0.5047, 0.4940, 0.4860, 0.4800, 0.4756, 0.4723])
finalReturns:  tensor([0.])
----------------------------------------
iter  1  stage  23  ep  99999   adversary:  AdversaryModes.fight_132
  actions:  tensor([8, 8, 0, 7, 7, 6, 1, 6, 1, 0, 7, 0, 7, 7, 0, 0, 9, 7, 8, 7,
0, 7, 7, 7,
        0])
loss=  tensor(0.0095, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.8497,
0.8497, 0.8497, 0.8497, 0.8497, 0.8497, 0.8497, 0.8497, 0.8497,
```

```
        0.8497, 0.8497, 0.8497, 0.8497, 0.8497, 0.8497, 0.8497, 0.8497, 0.8497,
        0.8497, 0.8497, 0.8497, 0.8497, 0.8497, 0.8497, 0.4157]) return=
112339.1163673332
probs of actions:  tensor([0.0652, 0.0703, 0.3638, 0.4614, 0.4650, 0.0731,
0.0180, 0.0588, 0.0144,
        0.3030, 0.4144, 0.2663, 0.4258, 0.3902, 0.3342, 0.2061, 0.0289, 0.4026,
        0.0797, 0.4148, 0.0888, 0.5708, 0.4182, 0.5905, 0.9986],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5048, 0.5468, 0.5857, 0.5202, 0.5014, 0.4889, 0.4787,
0.4481, 0.4484,
        0.4296, 0.4075, 0.4222, 0.4021, 0.4132, 0.4265, 0.4101, 0.3900, 0.4128,
        0.4198, 0.4311, 0.4400, 0.4152, 0.4231, 0.4291, 0.4385])
finalReturns:  tensor([0.0180, 0.0229])
----------------------------------------
iter  1  stage  22  ep  99999   adversary:  AdversaryModes.fight_132
  actions:  tensor([11, 18, 18, 18, 16, 18, 18, 14, 18, 18, 21, 15, 18, 18, 18,
11, 18, 19,
        18, 18, 11, 18, 18, 11,  0])
loss=  tensor(0.2019, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.3261,
1.3261, 1.3261, 1.3261, 1.3261, 1.3261, 1.3261, 1.3261, 1.3261,
        1.3261, 1.3261, 1.3261, 1.3261, 1.3261, 1.3261, 1.3261, 1.3261, 1.3261,
        1.3261, 1.3261, 1.3261, 1.3261, 1.3261, 0.8335, 0.3968]) return=
122146.33270204287
probs of actions:  tensor([0.1114, 0.5005, 0.4775, 0.4623, 0.0738, 0.4914,
0.4920, 0.0551, 0.5121,
        0.4809, 0.0076, 0.0585, 0.4930, 0.4906, 0.5019, 0.1138, 0.4775, 0.0571,
        0.4833, 0.4679, 0.1263, 0.5121, 0.6630, 0.2893, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5320, 0.5428, 0.5274, 0.5228, 0.5002, 0.4957,
0.5052, 0.4755,
        0.4772, 0.4668, 0.5002, 0.4775, 0.4788, 0.4797, 0.5006, 0.4561, 0.4589,
        0.4711, 0.4739, 0.4963, 0.4529, 0.4602, 0.4860, 0.4778])
finalReturns:  tensor([0.0979, 0.1303, 0.0809])
----------------------------------------
iter  1  stage  21  ep  99999   adversary:  AdversaryModes.fight_132
  actions:  tensor([18, 18, 18, 18, 19, 18, 17, 18, 18, 18, 16, 18, 18, 18, 18,
18, 18, 18,
        18, 16, 16, 18, 18, 18,  0])
loss=  tensor(0.0733, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.3379,
1.3379, 1.3379, 1.3379, 1.3379, 1.3379, 1.3379, 1.3379, 1.3379,
        1.3379, 1.3379, 1.3379, 1.3379, 1.3379, 1.3379, 1.3379, 1.3379, 1.3379,
        1.3379, 1.3379, 1.3379, 1.3379, 0.9393, 0.5929, 0.2833]) return=
102188.08265443008
probs of actions:  tensor([0.8722, 0.8706, 0.8676, 0.8419, 0.0511, 0.8602,
0.0068, 0.8830, 0.8689,
        0.8490, 0.0319, 0.8320, 0.8612, 0.8626, 0.8607, 0.8587, 0.8471, 0.8518,
        0.8542, 0.0333, 0.0318, 0.9187, 0.9270, 0.7905, 1.0000],
       grad_fn=<ExpBackward0>)
```

```
rewards:  tensor([0.4788, 0.5586, 0.5101, 0.4751, 0.4460, 0.4344, 0.4233,
0.4056, 0.3984,
        0.3931, 0.3959, 0.3796, 0.3790, 0.3786, 0.3782, 0.3780, 0.3778, 0.3776,
        0.3775, 0.3842, 0.3778, 0.3662, 0.3689, 0.3710, 0.4049])
finalReturns:  tensor([0.1732, 0.2056, 0.1830, 0.1216])
----------------------------------------
iter  1  stage  20  ep  99999   adversary:  AdversaryModes.fight_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0276, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.6350,
1.6350, 1.6350, 1.6350, 1.6350, 1.6350, 1.6350, 1.6350, 1.6350,
        1.6350, 1.6350, 1.6350, 1.6350, 1.6350, 1.6350, 1.6350, 1.6350, 1.6350,
        1.6350, 1.6350, 1.6350, 1.2248, 0.8702, 0.5549, 0.2674]) return=
102604.52812037413
probs of actions:  tensor([0.9710, 0.9696, 0.9702, 0.9607, 0.9690, 0.9665,
0.9702, 0.9737, 0.9688,
        0.9612, 0.9591, 0.9589, 0.9674, 0.9678, 0.9658, 0.9667, 0.9641, 0.9647,
        0.9651, 0.9602, 0.9635, 0.9843, 0.9838, 0.9696, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5101, 0.4751, 0.4497, 0.4310, 0.4172,
0.4071, 0.3995,
        0.3939, 0.3897, 0.3865, 0.3842, 0.3824, 0.3811, 0.3801, 0.3794, 0.3789,
        0.3784, 0.3781, 0.3779, 0.3777, 0.3776, 0.3775, 0.4098])
finalReturns:  tensor([0.2855, 0.3179, 0.2947, 0.2324, 0.1424])
----------------------------------------
iter  1  stage  19  ep  99999   adversary:  AdversaryModes.fight_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        16, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.0543, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.8695,
1.8695, 1.8695, 1.8695, 1.8695, 1.8695, 1.8695, 1.8695, 1.8695,
        1.8695, 1.8695, 1.8695, 1.8695, 1.8695, 1.8695, 1.8695, 1.8695, 1.8695,
        1.8695, 1.8695, 1.4654, 1.1152, 0.8028, 0.5175, 0.2517]) return=
102462.48329383403
probs of actions:  tensor([0.9692, 0.9680, 0.9683, 0.9593, 0.9679, 0.9651,
0.9692, 0.9716, 0.9667,
        0.9584, 0.9585, 0.9576, 0.9667, 0.9662, 0.9652, 0.9653, 0.9625, 0.9646,
        0.0057, 0.9604, 0.9625, 0.9790, 0.9841, 0.9763, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5101, 0.4751, 0.4497, 0.4310, 0.4172,
0.4071, 0.3995,
        0.3939, 0.3897, 0.3865, 0.3842, 0.3824, 0.3811, 0.3801, 0.3794, 0.3789,
        0.3852, 0.3717, 0.3731, 0.3741, 0.3749, 0.3755, 0.4083])
finalReturns:  tensor([0.4081, 0.4405, 0.4176, 0.3558, 0.2662, 0.1566])
----------------------------------------
iter  1  stage  18  ep  99999   adversary:  AdversaryModes.fight_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
```

```
18, 18, 18,
        18, 19, 18, 18, 18, 18,  0])
loss=  tensor(2.0158, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.1318,
2.1318, 2.1318, 2.1318, 2.1318, 2.1318, 2.1318, 2.1318, 2.1318,
        2.1318, 2.1318, 2.1318, 2.1318, 2.1318, 2.1318, 2.1318, 2.1318, 2.1318,
        2.1318, 1.7209, 1.3661, 1.0504, 0.7628, 0.4952, 0.2423]) return=
102665.3466452767
probs of actions:  tensor([0.9743, 0.9729, 0.9736, 0.9664, 0.9736, 0.9710,
0.9747, 0.9763, 0.9718,
        0.9644, 0.9661, 0.9645, 0.9721, 0.9718, 0.9708, 0.9715, 0.9687, 0.9713,
        0.9729, 0.0345, 0.9690, 0.9847, 0.9871, 0.9855, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5101, 0.4751, 0.4497, 0.4310, 0.4172,
0.4071, 0.3995,
        0.3939, 0.3897, 0.3865, 0.3842, 0.3824, 0.3811, 0.3801, 0.3794, 0.3789,
        0.3784, 0.3744, 0.3811, 0.3801, 0.3794, 0.3788, 0.4108])
finalReturns:  tensor([0.5514, 0.5838, 0.5642, 0.4988, 0.4063, 0.2945, 0.1686])
----------------------------------------
iter  1  stage  17  ep  99999   adversary:  AdversaryModes.fight_132
  actions:  tensor([18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18,
18, 18, 18,
        18, 18, 18, 18, 18, 18,  0])
loss=  tensor(0.2928, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.3675,
2.3675, 2.3675, 2.3675, 2.3675, 2.3675, 2.3675, 2.3675, 2.3675,
        2.3675, 2.3675, 2.3675, 2.3675, 2.3675, 2.3675, 2.3675, 2.3675, 2.3675,
        1.9563, 1.6011, 1.2853, 0.9975, 0.7298, 0.4768, 0.2344]) return=
102604.52812037413
probs of actions:  tensor([0.9312, 0.9290, 0.9287, 0.9162, 0.9312, 0.9262,
0.9345, 0.9344, 0.9254,
        0.9113, 0.9177, 0.9099, 0.9296, 0.9272, 0.9277, 0.9261, 0.9193, 0.9308,
        0.9225, 0.9046, 0.9087, 0.9552, 0.9603, 0.9625, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5586, 0.5101, 0.4751, 0.4497, 0.4310, 0.4172,
0.4071, 0.3995,
        0.3939, 0.3897, 0.3865, 0.3842, 0.3824, 0.3811, 0.3801, 0.3794, 0.3789,
        0.3784, 0.3781, 0.3779, 0.3777, 0.3776, 0.3775, 0.4098])
finalReturns:  tensor([0.6884, 0.7208, 0.6976, 0.6352, 0.5452, 0.4351, 0.3105,
0.1754])
----------------------------------------
iter  1  stage  16  ep  99999   adversary:  AdversaryModes.fight_132
  actions:  tensor([19, 18, 18, 19, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 19,
18, 18, 19,
        18, 19, 18, 19, 19, 18,  0])
loss=  tensor(4.2699, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.4183,
2.4183, 2.4183, 2.4183, 2.4183, 2.4183, 2.4183, 2.4183, 2.4183,
        2.4183, 2.4183, 2.4183, 2.4183, 2.4183, 2.4183, 2.4183, 2.4183, 2.0290,
        1.6949, 1.3995, 1.1316, 0.8834, 0.6494, 0.4260, 0.2101]) return=
97928.81160926864
```

```
probs of actions:  tensor([0.3446, 0.6577, 0.6444, 0.3681, 0.6642, 0.6525,
0.6800, 0.6510, 0.6417,
        0.6183, 0.6431, 0.6039, 0.6727, 0.6523, 0.3263, 0.6423, 0.6035, 0.3472,
        0.6241, 0.3860, 0.4948, 0.3028, 0.2832, 0.7703, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4751, 0.5624, 0.5055, 0.4610, 0.4386, 0.4161, 0.3996,
0.3875, 0.3785,
        0.3718, 0.3668, 0.3631, 0.3603, 0.3582, 0.3529, 0.3586, 0.3569, 0.3520,
        0.3579, 0.3527, 0.3584, 0.3531, 0.3550, 0.3602, 0.3905])
finalReturns:  tensor([0.8185, 0.8509, 0.8330, 0.7705, 0.6857, 0.5755, 0.4563,
0.3247, 0.1804])
----------------------------------------
iter  1  stage  15  ep  99999   adversary:  AdversaryModes.fight_132
  actions:  tensor([18, 19, 18, 19, 18, 18, 19, 18, 19, 19, 19, 18, 19, 18, 19,
18, 19, 19,
        18, 19, 19, 19, 19, 18,  0])
loss=  tensor(4.4872, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.8513,
2.8513, 2.8513, 2.8513, 2.8513, 2.8513, 2.8513, 2.8513, 2.8513, 2.8513,
        2.8513, 2.8513, 2.8513, 2.8513, 2.8513, 2.8513, 2.4309, 2.0693,
        1.7489, 1.4578, 1.1878, 0.9331, 0.6895, 0.4541, 0.2248]) return=
103665.53870452796
probs of actions:  tensor([0.2510, 0.7391, 0.2437, 0.7508, 0.2622, 0.2638,
0.7149, 0.2397, 0.7548,
        0.7597, 0.7351, 0.2216, 0.7160, 0.2554, 0.7117, 0.2246, 0.7894, 0.7670,
        0.2307, 0.7989, 0.8463, 0.7543, 0.7680, 0.3619, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5549, 0.5138, 0.4741, 0.4551, 0.4350, 0.4165,
0.4126, 0.3999,
        0.3965, 0.3940, 0.3958, 0.3874, 0.3908, 0.3837, 0.3881, 0.3816, 0.3828,
        0.3874, 0.3812, 0.3825, 0.3834, 0.3842, 0.3884, 0.4180])
finalReturns:  tensor([1.0263, 1.0587, 1.0387, 0.9762, 0.8799, 0.7687, 0.6410,
0.5011, 0.3523,
        0.1932])
----------------------------------------
iter  1  stage  14  ep  99999   adversary:  AdversaryModes.fight_132
  actions:  tensor([18, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19,
18, 19, 19,
        19, 19, 19, 19, 19, 19,  0])
loss=  tensor(3.7333, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.0920,
3.0920, 3.0920, 3.0920, 3.0920, 3.0920, 3.0920, 3.0920, 3.0920, 3.0920,
        3.0920, 3.0920, 3.0920, 3.0920, 3.0920, 3.0920, 2.6659, 2.3003, 1.9772,
        1.6841, 1.4127, 1.1569, 0.9125, 0.6766, 0.4468, 0.2217]) return=
104293.09788509886
probs of actions:  tensor([0.0991, 0.8924, 0.9044, 0.8958, 0.8924, 0.8902,
0.8814, 0.9084, 0.9020,
        0.9004, 0.8872, 0.9107, 0.8797, 0.8967, 0.8710, 0.0876, 0.9185, 0.9141,
        0.9163, 0.9272, 0.9501, 0.9235, 0.9188, 0.8713, 1.0000],
       grad_fn=<ExpBackward0>)
```

```
rewards:  tensor([0.4788, 0.5549, 0.5101, 0.4777, 0.4540, 0.4367, 0.4238,
0.4143, 0.4073,
        0.4020, 0.3981, 0.3951, 0.3929, 0.3913, 0.3901, 0.3929, 0.3852, 0.3855,
        0.3857, 0.3859, 0.3860, 0.3861, 0.3862, 0.3862, 0.4224])
finalReturns:  tensor([1.2002, 1.2363, 1.2090, 1.1469, 1.0545, 0.9402, 0.8101,
0.6684, 0.5182,
        0.3618, 0.2007])
----------------------------------------
iter  1  stage  13  ep  99999   adversary:  AdversaryModes.fight_132
  actions:  tensor([19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19,
19, 19, 19,
        19, 19, 19, 19, 19, 19,  0])
loss=  tensor(0.1330, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.0724,
3.0724, 3.0724, 3.0724, 3.0724, 3.0724, 3.0724, 3.0724, 3.0724,
        3.0724, 3.0724, 3.0724, 3.0724, 3.0724, 2.6699, 2.3266, 2.0248, 1.7523,
        1.5007, 1.2644, 1.0391, 0.8220, 0.6109, 0.4043, 0.2009]) return=
99272.14211264676
probs of actions:  tensor([0.9851, 0.9828, 0.9857, 0.9825, 0.9828, 0.9817,
0.9803, 0.9868, 0.9846,
        0.9830, 0.9800, 0.9857, 0.9797, 0.9868, 0.9807, 0.9862, 0.9914, 0.9871,
        0.9861, 0.9902, 0.9942, 0.9929, 0.9898, 0.9844, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4751, 0.5587, 0.5055, 0.4672, 0.4394, 0.4191, 0.4042,
0.3931, 0.3849,
        0.3788, 0.3743, 0.3709, 0.3684, 0.3665, 0.3650, 0.3640, 0.3632, 0.3626,
        0.3621, 0.3618, 0.3616, 0.3614, 0.3612, 0.3611, 0.3971])
finalReturns:  tensor([1.3152, 1.3513, 1.3294, 1.2673, 1.1766, 1.0656, 0.9398,
0.8033, 0.6588,
        0.5086, 0.3540, 0.1962])
----------------------------------------
iter  1  stage  12  ep  99999   adversary:  AdversaryModes.fight_132
  actions:  tensor([19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19,
19, 19, 19,
        19, 19, 19, 19, 19, 19,  0])
loss=  tensor(0.0689, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.2779,
3.2779, 3.2779, 3.2779, 3.2779, 3.2779, 3.2779, 3.2779, 3.2779,
        3.2779, 3.2779, 3.2779, 3.2779, 2.8735, 2.5289, 2.2262, 1.9530, 1.7010,
        1.4643, 1.2388, 1.0215, 0.8103, 0.6035, 0.4001, 0.1991]) return=
99272.14211264676
probs of actions:  tensor([0.9934, 0.9920, 0.9937, 0.9918, 0.9921, 0.9914,
0.9908, 0.9942, 0.9929,
        0.9919, 0.9907, 0.9933, 0.9896, 0.9951, 0.9918, 0.9950, 0.9964, 0.9951,
        0.9949, 0.9964, 0.9983, 0.9980, 0.9968, 0.9942, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4751, 0.5587, 0.5055, 0.4672, 0.4394, 0.4191, 0.4042,
0.3931, 0.3849,
        0.3788, 0.3743, 0.3709, 0.3684, 0.3665, 0.3650, 0.3640, 0.3632, 0.3626,
        0.3621, 0.3618, 0.3616, 0.3614, 0.3612, 0.3611, 0.3971])
```

```
finalReturns:  tensor([1.4780, 1.5141, 1.4922, 1.4299, 1.3391, 1.2279, 1.1020,
0.9654, 0.8209,
        0.6706, 0.5160, 0.3582, 0.1980])
---------------------------------------
iter  1  stage  11  ep  99999   adversary:  AdversaryModes.fight_132
  actions:  tensor([19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19,
19, 19, 19,
        19, 19, 19, 19, 19, 19,  0])
loss=  tensor(0.0430, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.4843,
3.4843, 3.4843, 3.4843, 3.4843, 3.4843, 3.4843, 3.4843, 3.4843,
        3.4843, 3.4843, 3.4843, 3.0773, 2.7310, 2.4271, 2.1530, 1.9004, 1.6632,
        1.4374, 1.2198, 1.0084, 0.8015, 0.5979, 0.3969, 0.1978]) return=
99272.14211264676
probs of actions:  tensor([0.9958, 0.9948, 0.9961, 0.9946, 0.9950, 0.9944,
0.9941, 0.9963, 0.9954,
        0.9947, 0.9939, 0.9959, 0.9941, 0.9979, 0.9955, 0.9974, 0.9982, 0.9979,
        0.9976, 0.9987, 0.9994, 0.9993, 0.9986, 0.9973, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4751, 0.5587, 0.5055, 0.4672, 0.4394, 0.4191, 0.4042,
0.3931, 0.3849,
        0.3788, 0.3743, 0.3709, 0.3684, 0.3665, 0.3650, 0.3640, 0.3632, 0.3626,
        0.3621, 0.3618, 0.3616, 0.3614, 0.3612, 0.3611, 0.3971])
finalReturns:  tensor([1.6425, 1.6786, 1.6565, 1.5941, 1.5031, 1.3918, 1.2657,
1.1290, 0.9844,
        0.8340, 0.6794, 0.5215, 0.3613, 0.1994])
---------------------------------------
iter  1  stage  10  ep  96876   adversary:  AdversaryModes.fight_132
  actions:  tensor([18, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19,
19, 19, 19,
        19, 19, 19, 19, 19, 19,  0])
loss=  tensor(0.0057, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.9867,
3.9867, 3.9867, 3.9867, 3.9867, 3.9867, 3.9867, 3.9867, 3.9867,
        3.9867, 3.9867, 3.5525, 3.1814, 2.8543, 2.5585, 2.2850, 2.0277, 1.7823,
        1.5456, 1.3153, 1.0897, 0.8677, 0.6483, 0.4308, 0.2149]) return=
104376.37516292125
probs of actions:  tensor([8.0187e-04, 9.9891e-01, 9.9926e-01, 9.9889e-01,
9.9902e-01, 9.9886e-01,
        9.9884e-01, 9.9934e-01, 9.9911e-01, 9.9894e-01, 9.9903e-01, 9.9962e-01,
        9.9943e-01, 1.0000e+00, 9.9945e-01, 9.9960e-01, 9.9995e-01, 9.9999e-01,
        9.9977e-01, 1.0000e+00, 1.0000e+00, 1.0000e+00, 1.0000e+00, 9.9992e-01,
        1.0000e+00], grad_fn=<ExpBackward0>)
rewards:  tensor([0.4788, 0.5549, 0.5101, 0.4777, 0.4540, 0.4367, 0.4238,
0.4143, 0.4073,
        0.4020, 0.3981, 0.3951, 0.3929, 0.3913, 0.3901, 0.3892, 0.3885, 0.3879,
        0.3876, 0.3873, 0.3871, 0.3869, 0.3868, 0.3867, 0.4227])
finalReturns:  tensor([1.8914, 1.9275, 1.9034, 1.8376, 1.7421, 1.6255, 1.4936,
1.3505, 1.1993,
        1.0421, 0.8804, 0.7154, 0.5479, 0.3785, 0.2078])
```

```
---------------------------------------
iter  1  stage  9  ep  64   adversary:  AdversaryModes.fight_132
  actions:  tensor([19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19,
19, 19, 19,
        19, 19, 19, 19, 19, 19,  0])
loss=  tensor(0.0079, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.9042,
3.9042, 3.9042, 3.9042, 3.9042, 3.9042, 3.9042, 3.9042, 3.9042,
        3.9042, 3.4893, 3.1375, 2.8297, 2.5529, 2.2983, 2.0597, 1.8328, 1.6145,
        1.4025, 1.1952, 0.9913, 0.7900, 0.5907, 0.3928, 0.1960]) return=
99272.14211264676
probs of actions:  tensor([0.9992, 0.9989, 0.9993, 0.9989, 0.9990, 0.9989,
0.9988, 0.9993, 0.9991,
        0.9990, 0.9990, 0.9996, 0.9994, 1.0000, 0.9994, 0.9996, 0.9999, 1.0000,
        0.9998, 1.0000, 1.0000, 1.0000, 1.0000, 0.9999, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4751, 0.5587, 0.5055, 0.4672, 0.4394, 0.4191, 0.4042,
0.3931, 0.3849,
        0.3788, 0.3743, 0.3709, 0.3684, 0.3665, 0.3650, 0.3640, 0.3632, 0.3626,
        0.3621, 0.3618, 0.3616, 0.3614, 0.3612, 0.3611, 0.3971])
finalReturns:  tensor([1.9757, 2.0118, 1.9893, 1.9262, 1.8347, 1.7229, 1.5964,
1.4593, 1.3144,
        1.1639, 1.0090, 0.8511, 0.6908, 0.5288, 0.3654, 0.2011])
---------------------------------------
iter  1  stage  8  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19,
19, 19, 19,
        19, 19, 19, 19, 19, 19,  0])
loss=  tensor(0.0104, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.1205,
4.1205, 4.1205, 4.1205, 4.1205, 4.1205, 4.1205, 4.1205, 4.1205,
        3.6995, 3.3435, 3.0328, 2.7539, 2.4977, 2.2581, 2.0304, 1.8115, 1.5990,
        1.3914, 1.1873, 0.9859, 0.7864, 0.5884, 0.3915, 0.1954]) return=
99272.14211264676
probs of actions:  tensor([0.9992, 0.9989, 0.9993, 0.9989, 0.9990, 0.9989,
0.9988, 0.9993, 0.9991,
        0.9990, 0.9990, 0.9996, 0.9994, 1.0000, 0.9994, 0.9996, 0.9999, 1.0000,
        0.9998, 1.0000, 1.0000, 1.0000, 1.0000, 0.9999, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4751, 0.5587, 0.5055, 0.4672, 0.4394, 0.4191, 0.4042,
0.3931, 0.3849,
        0.3788, 0.3743, 0.3709, 0.3684, 0.3665, 0.3650, 0.3640, 0.3632, 0.3626,
        0.3621, 0.3618, 0.3616, 0.3614, 0.3612, 0.3611, 0.3971])
finalReturns:  tensor([2.1443, 2.1804, 2.1575, 2.0941, 2.0021, 1.8898, 1.7631,
1.6257, 1.4806,
        1.3299, 1.1750, 1.0169, 0.8565, 0.6945, 0.5311, 0.3667, 0.2017])
---------------------------------------
iter  1  stage  7  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19,
19, 19, 19,
```

```
        19, 19, 19, 19, 19, 19,  0])
loss=  tensor(0.0124, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.3436,
4.3436, 4.3436, 4.3436, 4.3436, 4.3436, 4.3436, 4.3436, 3.9144,
        3.5528, 3.2380, 2.9563, 2.6982, 2.4571, 2.2283, 2.0086, 1.7956, 1.5875,
        1.3831, 1.1814, 0.9818, 0.7837, 0.5867, 0.3905, 0.1950]) return=
99272.14211264676
probs of actions:  tensor([0.9992, 0.9989, 0.9993, 0.9989, 0.9990, 0.9989,
0.9988, 0.9993, 0.9991,
        0.9990, 0.9990, 0.9996, 0.9994, 1.0000, 0.9994, 0.9996, 0.9999, 1.0000,
        0.9998, 1.0000, 1.0000, 1.0000, 1.0000, 0.9999, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4751, 0.5587, 0.5055, 0.4672, 0.4394, 0.4191, 0.4042,
0.3931, 0.3849,
        0.3788, 0.3743, 0.3709, 0.3684, 0.3665, 0.3650, 0.3640, 0.3632, 0.3626,
        0.3621, 0.3618, 0.3616, 0.3614, 0.3612, 0.3611, 0.3971])
finalReturns:  tensor([2.3143, 2.3504, 2.3271, 2.2631, 2.1705, 2.0577, 1.9305,
1.7928, 1.6475,
        1.4965, 1.3414, 1.1832, 1.0228, 0.8606, 0.6972, 0.5328, 0.3677, 0.2021])
----------------------------------------
iter  1  stage  6  ep  98   adversary:  AdversaryModes.fight_132
  actions:  tensor([19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19,
19, 19, 19,
        19, 19, 19, 19, 19, 19,  0])
loss=  tensor(0.0142, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.5761,
4.5761, 4.5761, 4.5761, 4.5761, 4.5761, 4.5761, 4.1358, 3.7666,
        3.4465, 3.1610, 2.9002, 2.6571, 2.4269, 2.2062, 1.9924, 1.7837, 1.5789,
        1.3769, 1.1770, 0.9787, 0.7816, 0.5854, 0.3898, 0.1947]) return=
99272.14211264676
probs of actions:  tensor([0.9993, 0.9990, 0.9993, 0.9990, 0.9991, 0.9989,
0.9990, 0.9994, 0.9992,
        0.9991, 0.9991, 0.9996, 0.9994, 1.0000, 0.9995, 0.9996, 0.9999, 1.0000,
        0.9998, 1.0000, 1.0000, 1.0000, 1.0000, 0.9999, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4751, 0.5587, 0.5055, 0.4672, 0.4394, 0.4191, 0.4042,
0.3931, 0.3849,
        0.3788, 0.3743, 0.3709, 0.3684, 0.3665, 0.3650, 0.3640, 0.3632, 0.3626,
        0.3621, 0.3618, 0.3616, 0.3614, 0.3612, 0.3611, 0.3971])
finalReturns:  tensor([2.4860, 2.5221, 2.4983, 2.4334, 2.3400, 2.2266, 2.0988,
1.9607, 1.8149,
        1.6637, 1.5084, 1.3500, 1.1894, 1.0272, 0.8637, 0.6992, 0.5341, 0.3685,
        0.2024])
----------------------------------------
iter  1  stage  5  ep  43   adversary:  AdversaryModes.fight_132
  actions:  tensor([19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19,
19, 19, 19,
        19, 19, 19, 19, 19, 19,  0])
loss=  tensor(0.0167, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.8214,
4.8214, 4.8214, 4.8214, 4.8214, 4.8214, 4.3662, 3.9867, 3.6595,
```

```
        3.3690, 3.1045, 2.8588, 2.6267, 2.4045, 2.1897, 1.9803, 1.7749, 1.5725,
        1.3723, 1.1737, 0.9764, 0.7801, 0.5844, 0.3892, 0.1945]) return=
99272.14211264676
probs of actions:  tensor([0.9993, 0.9990, 0.9993, 0.9990, 0.9991, 0.9990,
0.9991, 0.9995, 0.9993,
        0.9992, 0.9991, 0.9996, 0.9995, 1.0000, 0.9995, 0.9996, 0.9999, 1.0000,
        0.9998, 1.0000, 1.0000, 1.0000, 1.0000, 0.9999, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4751, 0.5587, 0.5055, 0.4672, 0.4394, 0.4191, 0.4042,
0.3931, 0.3849,
        0.3788, 0.3743, 0.3709, 0.3684, 0.3665, 0.3650, 0.3640, 0.3632, 0.3626,
        0.3621, 0.3618, 0.3616, 0.3614, 0.3612, 0.3611, 0.3971])
finalReturns:  tensor([2.6598, 2.6959, 2.6712, 2.6053, 2.5109, 2.3966, 2.2680,
2.1292, 1.9830,
        1.8314, 1.6758, 1.5172, 1.3565, 1.1941, 1.0305, 0.8660, 0.7008, 0.5351,
        0.3690, 0.2027])
----------------------------------------
iter  1  stage  4  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19,
19, 19, 19,
        19, 19, 19, 19, 19, 19,  0])
loss=  tensor(0.0199, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.0845,
5.0845, 5.0845, 5.0845, 5.0845, 4.6089, 4.2155, 3.8787, 3.5813,
        3.3120, 3.0628, 2.8281, 2.6041, 2.3879, 2.1774, 1.9713, 1.7683, 1.5677,
        1.3688, 1.1713, 0.9747, 0.7789, 0.5837, 0.3888, 0.1943]) return=
99272.14211264676
probs of actions:  tensor([0.9993, 0.9990, 0.9993, 0.9990, 0.9991, 0.9990,
0.9991, 0.9995, 0.9993,
        0.9992, 0.9991, 0.9996, 0.9995, 1.0000, 0.9995, 0.9996, 0.9999, 1.0000,
        0.9998, 1.0000, 1.0000, 1.0000, 1.0000, 0.9999, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4751, 0.5587, 0.5055, 0.4672, 0.4394, 0.4191, 0.4042,
0.3931, 0.3849,
        0.3788, 0.3743, 0.3709, 0.3684, 0.3665, 0.3650, 0.3640, 0.3632, 0.3626,
        0.3621, 0.3618, 0.3616, 0.3614, 0.3612, 0.3611, 0.3971])
finalReturns:  tensor([2.8362, 2.8723, 2.8465, 2.7792, 2.6835, 2.5679, 2.4383,
2.2987, 2.1518,
        1.9997, 1.8437, 1.6848, 1.5238, 1.3613, 1.1975, 1.0329, 0.8677, 0.7019,
        0.5358, 0.3694, 0.2028])
----------------------------------------
iter  1  stage  3  ep  10   adversary:  AdversaryModes.fight_132
  actions:  tensor([19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19,
19, 19, 19,
        19, 19, 19, 19, 19, 19,  0])
loss=  tensor(0.0232, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.3719,
5.3719, 5.3719, 5.3719, 4.8686, 4.4563, 4.1064, 3.7998, 3.5239,
        3.2700, 3.0319, 2.8053, 2.5873, 2.3754, 2.1683, 1.9645, 1.7633, 1.5641,
        1.3662, 1.1694, 0.9734, 0.7781, 0.5831, 0.3885, 0.1942]) return=
```

```
99272.14211264676
probs of actions:  tensor([0.9993, 0.9990, 0.9993, 0.9990, 0.9991, 0.9990,
0.9991, 0.9995, 0.9993,
        0.9992, 0.9991, 0.9996, 0.9995, 1.0000, 0.9995, 0.9996, 0.9999, 1.0000,
        0.9998, 1.0000, 1.0000, 1.0000, 1.0000, 0.9999, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4751, 0.5587, 0.5055, 0.4672, 0.4394, 0.4191, 0.4042,
0.3931, 0.3849,
        0.3788, 0.3743, 0.3709, 0.3684, 0.3665, 0.3650, 0.3640, 0.3632, 0.3626,
        0.3621, 0.3618, 0.3616, 0.3614, 0.3612, 0.3611, 0.3971])
finalReturns:  tensor([3.0159, 3.0520, 3.0249, 2.9557, 2.8581, 2.7409, 2.6099,
2.4692, 2.3215,
        2.1687, 2.0121, 1.8529, 1.6916, 1.5288, 1.3649, 1.2001, 1.0348, 0.8690,
        0.7028, 0.5364, 0.3697, 0.2030])
----------------------------------------
iter  1  stage  2  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19,
19, 19, 19,
        19, 19, 19, 19, 19, 19,  0])
loss=  tensor(0.0260, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.6933,
5.6933, 5.6933, 5.1517, 4.7135, 4.3457, 4.0266, 3.7419, 3.4815,
        3.2388, 3.0089, 2.7884, 2.5747, 2.3662, 2.1614, 1.9595, 1.7596, 1.5614,
        1.3643, 1.1680, 0.9725, 0.7774, 0.5827, 0.3883, 0.1941]) return=
99272.14211264676
probs of actions:  tensor([0.9993, 0.9990, 0.9993, 0.9990, 0.9991, 0.9990,
0.9991, 0.9995, 0.9993,
        0.9992, 0.9991, 0.9996, 0.9995, 1.0000, 0.9995, 0.9996, 0.9999, 1.0000,
        0.9998, 1.0000, 1.0000, 1.0000, 1.0000, 0.9999, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4751, 0.5587, 0.5055, 0.4672, 0.4394, 0.4191, 0.4042,
0.3931, 0.3849,
        0.3788, 0.3743, 0.3709, 0.3684, 0.3665, 0.3650, 0.3640, 0.3632, 0.3626,
        0.3621, 0.3618, 0.3616, 0.3614, 0.3612, 0.3611, 0.3971])
finalReturns:  tensor([3.2001, 3.2362, 3.2071, 3.1355, 3.0354, 2.9161, 2.7833,
2.6411, 2.4922,
        2.3385, 2.1812, 2.0214, 1.8597, 1.6966, 1.5325, 1.3676, 1.2021, 1.0362,
        0.8699, 0.7034, 0.5368, 0.3700, 0.2031])
----------------------------------------
iter  1  stage  1  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19,
19, 19, 19,
        19, 19, 19, 19, 19, 19,  0])
loss=  tensor(0.0299, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.0620,
6.0620, 5.4672, 4.9933, 4.6010, 4.2649, 3.9681, 3.6991, 3.4501,
        3.2157, 2.9918, 2.7757, 2.5653, 2.3592, 2.1563, 1.9557, 1.7569, 1.5593,
        1.3628, 1.1670, 0.9718, 0.7769, 0.5824, 0.3881, 0.1940]) return=
99272.14211264676
probs of actions:  tensor([0.9993, 0.9990, 0.9993, 0.9990, 0.9991, 0.9990,
```

```
0.9991, 0.9995, 0.9993,
        0.9992, 0.9991, 0.9996, 0.9995, 1.0000, 0.9995, 0.9996, 0.9999, 1.0000,
        0.9998, 1.0000, 1.0000, 1.0000, 1.0000, 0.9999, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4751, 0.5587, 0.5055, 0.4672, 0.4394, 0.4191, 0.4042,
0.3931, 0.3849,
        0.3788, 0.3743, 0.3709, 0.3684, 0.3665, 0.3650, 0.3640, 0.3632, 0.3626,
        0.3621, 0.3618, 0.3616, 0.3614, 0.3612, 0.3611, 0.3971])
finalReturns:  tensor([3.3901, 3.4262, 3.3946, 3.3197, 3.2163, 3.0940, 2.9588,
2.8147, 2.6642,
        2.5093, 2.3511, 2.1906, 2.0284, 1.8649, 1.7004, 1.5353, 1.3696, 1.2035,
        1.0372, 0.8706, 0.7039, 0.5371, 0.3701, 0.2031])
----------------------------------------
iter  1  stage  0  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19,
19, 19, 19,
        19, 19, 19, 19, 19, 19,  0])
loss=  tensor(0.0331, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.3395,
5.8282, 5.3044, 4.8783, 4.5187, 4.2054, 3.9248, 3.6674, 3.4268,
        3.1984, 2.9790, 2.7662, 2.5583, 2.3540, 2.1524, 1.9529, 1.7548, 1.5578,
        1.3617, 1.1662, 0.9712, 0.7766, 0.5822, 0.3880, 0.1939]) return=
99272.14211264676
probs of actions:  tensor([0.9993, 0.9990, 0.9993, 0.9990, 0.9991, 0.9990,
0.9991, 0.9995, 0.9993,
        0.9992, 0.9991, 0.9996, 0.9995, 1.0000, 0.9995, 0.9996, 0.9999, 1.0000,
        0.9998, 1.0000, 1.0000, 1.0000, 1.0000, 0.9999, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4751, 0.5587, 0.5055, 0.4672, 0.4394, 0.4191, 0.4042,
0.3931, 0.3849,
        0.3788, 0.3743, 0.3709, 0.3684, 0.3665, 0.3650, 0.3640, 0.3632, 0.3626,
        0.3621, 0.3618, 0.3616, 0.3614, 0.3612, 0.3611, 0.3971])
finalReturns:  tensor([3.5878, 3.6239, 3.5889, 3.5096, 3.4019, 3.2758, 3.1373,
2.9905, 2.8380,
        2.6815, 2.5221, 2.3606, 2.1976, 2.0336, 1.8687, 1.7032, 1.5373, 1.3711,
        1.2046, 1.0380, 0.8712, 0.7043, 0.5373, 0.3703, 0.2032])
0,[1e-05,1][1, 10000, 1, 1],1682974149 saved
[1777102, 'tensor([0., 0., 0., 0., 0., 1., 0., 0., 0., 0., 0., 0.])',
99272.14211264676, 86128.77669412928, 0.033126410096883774, 1e-05, 1, 0,
'tensor([19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19,
19,\n        19, 19, 19, 19, 19, 19,  0])', '[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.\n 1.]', '0,[1e-05,1][1, 10000, 1,
1],1682974149', 25, 50, 165673.58346428472, 196485.87523713993,
79648.59533906801, 135210.39999999997, 132281.41866666666, 99272.14211264675,
99263.9372312988, 117110.65916440394, 117121.26544005591, 92350.73961268678,
99272.14211264675, 117121.26544005591]
policy reset
----------------------------------------
iter  2  stage  24  ep  99999   adversary:  AdversaryModes.fight_132
```

```
  actions:  tensor([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0,
        0])
loss=  tensor(-0., grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.5624,
0.5624, 0.5624, 0.5624, 0.5624, 0.5624, 0.5624, 0.5624, 0.5624,
        0.5624, 0.5624, 0.5624, 0.5624, 0.5624, 0.5624, 0.5624, 0.5624, 0.5624,
        0.5624, 0.5624, 0.5624, 0.5624, 0.5624, 0.5624, 0.5624]) return=
138554.5803254051
probs of actions:  tensor([0.8737, 0.8892, 0.8547, 0.8885, 0.9043, 0.8971,
0.8719, 0.9021, 0.9021,
        0.8920, 0.8935, 0.8782, 0.8894, 0.8990, 0.8990, 0.9052, 0.8961, 0.8807,
        0.8858, 0.8928, 0.8808, 0.9072, 0.9093, 0.9040, 0.9838],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5112, 0.5238, 0.5334, 0.5406, 0.5460, 0.5501, 0.5532,
0.5555, 0.5573,
        0.5586, 0.5595, 0.5603, 0.5608, 0.5613, 0.5616, 0.5618, 0.5620, 0.5621,
        0.5622, 0.5623, 0.5623, 0.5624, 0.5624, 0.5624, 0.5624])
finalReturns:  tensor([0.])
----------------------------------------
iter  2  stage  23  ep  99999   adversary:  AdversaryModes.fight_132
  actions:  tensor([11, 11, 11, 13, 11,  0,  0,  0,  7, 11, 16,  7, 11,  7,  1,
8,  5,  7,
         1,  5,  6, 11, 13, 11,  0])
loss=  tensor(0.0061, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.9062,
0.9062, 0.9062, 0.9062, 0.9062, 0.9062, 0.9062, 0.9062, 0.9062,
        0.9062, 0.9062, 0.9062, 0.9062, 0.9062, 0.9062, 0.9062, 0.9062, 0.9062,
        0.9062, 0.9062, 0.9062, 0.9062, 0.9062, 0.9062, 0.4390]) return=
116362.73893427021
probs of actions:  tensor([4.2262e-01, 4.3894e-01, 4.4018e-01, 3.8976e-03,
4.5545e-01, 1.8125e-01,
        1.2230e-01, 1.1571e-01, 4.2787e-02, 4.4436e-01, 9.1077e-04, 5.0764e-02,
        5.1019e-01, 4.6578e-02, 3.3308e-02, 1.2511e-01, 3.2285e-02, 4.3355e-02,
        2.2989e-02, 2.9944e-02, 5.5487e-02, 4.7207e-01, 2.5546e-03, 7.9477e-01,
        9.9147e-01], grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5207, 0.5243, 0.5282, 0.4832,
0.4507, 0.4222,
        0.4204, 0.4243, 0.4754, 0.4603, 0.4753, 0.4722, 0.4397, 0.4477, 0.4383,
        0.4445, 0.4234, 0.4214, 0.4156, 0.4293, 0.4550, 0.4762])
finalReturns:  tensor([0.0251, 0.0372])
----------------------------------------
iter  2  stage  22  ep  89767   adversary:  AdversaryModes.fight_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11, 11, 11,
        11, 11, 11, 11, 11, 11,  0])
loss=  tensor(0.0003, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.4083,
1.4083, 1.4083, 1.4083, 1.4083, 1.4083, 1.4083, 1.4083, 1.4083,
        1.4083, 1.4083, 1.4083, 1.4083, 1.4083, 1.4083, 1.4083, 1.4083, 1.4083,
        1.4083, 1.4083, 1.4083, 1.4083, 1.4083, 0.9041, 0.4382]) return=
```

```
125571.79187222246
probs of actions:  tensor([0.9818, 0.9901, 0.9821, 0.9903, 0.9905, 0.9868,
0.9907, 0.9941, 0.9928,
        0.9893, 0.9932, 0.9848, 0.9921, 0.9868, 0.9691, 0.9914, 0.9843, 0.9874,
        0.9920, 0.9875, 0.9880, 0.9906, 0.9990, 0.9993, 0.9979],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4923, 0.4922, 0.4922, 0.4921, 0.4921, 0.5042])
finalReturns:  tensor([0.0801, 0.0922, 0.0660])
----------------------------------------
iter  2  stage  21  ep  39489   adversary:  AdversaryModes.fight_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11, 11, 11,
        11, 11, 11, 11, 11, 11,  0])
loss=  tensor(0.0004, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.8263,
1.8263, 1.8263, 1.8263, 1.8263, 1.8263, 1.8263, 1.8263, 1.8263, 1.8263,
        1.8263, 1.8263, 1.8263, 1.8263, 1.8263, 1.8263, 1.8263, 1.8263, 1.8263,
        1.8263, 1.8263, 1.8263, 1.8263, 1.8263, 1.3220, 0.8561, 0.4179]) return=
125571.79187222246
probs of actions:  tensor([0.9967, 0.9983, 0.9966, 0.9982, 0.9983, 0.9977,
0.9982, 0.9989, 0.9987,
        0.9981, 0.9987, 0.9971, 0.9985, 0.9973, 0.9934, 0.9985, 0.9974, 0.9975,
        0.9984, 0.9977, 0.9977, 0.9990, 0.9999, 0.9999, 0.9970],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4923, 0.4922, 0.4922, 0.4921, 0.4921, 0.5042])
finalReturns:  tensor([0.1543, 0.1664, 0.1402, 0.0863])
----------------------------------------
iter  2  stage  20  ep  15903   adversary:  AdversaryModes.fight_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11, 11, 11,
        11, 11, 11, 11, 11, 11,  0])
loss=  tensor(0.0008, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.2295,
2.2295, 2.2295, 2.2295, 2.2295, 2.2295, 2.2295, 2.2295, 2.2295, 2.2295,
        2.2295, 2.2295, 2.2295, 2.2295, 2.2295, 2.2295, 2.2295, 2.2295, 2.2295,
        2.2295, 2.2295, 2.2295, 1.7252, 1.2592, 0.8210, 0.4030]) return=
125571.79187222246
probs of actions:  tensor([0.9983, 0.9991, 0.9982, 0.9991, 0.9991, 0.9988,
0.9990, 0.9994, 0.9993,
        0.9990, 0.9993, 0.9984, 0.9992, 0.9985, 0.9964, 0.9992, 0.9986, 0.9987,
        0.9991, 0.9987, 0.9990, 0.9995, 0.9999, 1.0000, 0.9962],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
```

```
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4923, 0.4922, 0.4922, 0.4921, 0.4921, 0.5042])
finalReturns:  tensor([0.2433, 0.2554, 0.2292, 0.1753, 0.1011])
----------------------------------------
iter  2  stage  19  ep  94   adversary:  AdversaryModes.fight_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11, 11, 11,
        11, 11, 11, 11, 11, 11,  0])
loss=  tensor(0.0013, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.6218,
2.6218, 2.6218, 2.6218, 2.6218, 2.6218, 2.6218, 2.6218, 2.6218,
        2.6218, 2.6218, 2.6218, 2.6218, 2.6218, 2.6218, 2.6218, 2.6218, 2.6218,
        2.6218, 2.6218, 2.1174, 1.6514, 1.2131, 0.7951, 0.3921]) return=
125571.79187222246
probs of actions:  tensor([0.9983, 0.9991, 0.9982, 0.9991, 0.9991, 0.9988,
0.9991, 0.9995, 0.9993,
        0.9990, 0.9993, 0.9985, 0.9992, 0.9986, 0.9965, 0.9992, 0.9987, 0.9987,
        0.9992, 0.9990, 0.9990, 0.9995, 0.9999, 1.0000, 0.9961],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4923, 0.4922, 0.4922, 0.4921, 0.4921, 0.5042])
finalReturns:  tensor([0.3433, 0.3554, 0.3292, 0.2753, 0.2011, 0.1121])
----------------------------------------
iter  2  stage  18  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11, 11, 11,
        11, 11, 11, 11, 11, 11,  0])
loss=  tensor(0.0019, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.0061,
3.0061, 3.0061, 3.0061, 3.0061, 3.0061, 3.0061, 3.0061, 3.0061,
        3.0061, 3.0061, 3.0061, 3.0061, 3.0061, 3.0061, 3.0061, 3.0061, 3.0061,
        3.0061, 2.5016, 2.0355, 1.5972, 1.1791, 0.7760, 0.3839]) return=
125571.79187222246
probs of actions:  tensor([0.9983, 0.9991, 0.9982, 0.9991, 0.9991, 0.9988,
0.9991, 0.9995, 0.9993,
        0.9990, 0.9993, 0.9985, 0.9992, 0.9986, 0.9965, 0.9992, 0.9987, 0.9987,
        0.9992, 0.9990, 0.9990, 0.9995, 0.9999, 1.0000, 0.9961],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4923, 0.4922, 0.4922, 0.4921, 0.4921, 0.5042])
finalReturns:  tensor([0.4515, 0.4636, 0.4374, 0.3835, 0.3093, 0.2202, 0.1202])
----------------------------------------
iter  2  stage  17  ep  155   adversary:  AdversaryModes.fight_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11, 11, 11,
        11, 11, 11, 11, 11, 11,  0])
```

```
loss=  tensor(0.0026, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.3845,
3.3845, 3.3845, 3.3845, 3.3845, 3.3845, 3.3845, 3.3845, 3.3845,
        3.3845, 3.3845, 3.3845, 3.3845, 3.3845, 3.3845, 3.3845, 3.3845, 3.3845,
        2.8798, 2.4136, 1.9752, 1.5571, 1.1540, 0.7619, 0.3779]) return=
125571.79187222246
probs of actions:  tensor([0.9983, 0.9992, 0.9983, 0.9991, 0.9992, 0.9988,
0.9991, 0.9995, 0.9993,
        0.9991, 0.9994, 0.9985, 0.9992, 0.9986, 0.9965, 0.9992, 0.9987, 0.9990,
        0.9994, 0.9990, 0.9991, 0.9996, 0.9999, 1.0000, 0.9964],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4923, 0.4922, 0.4922, 0.4921, 0.4921, 0.5042])
finalReturns:  tensor([0.5657, 0.5778, 0.5516, 0.4976, 0.4235, 0.3344, 0.2344,
0.1263])
----------------------------------------
iter  2  stage  16  ep  95   adversary:  AdversaryModes.fight_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11, 11, 11,
        11, 11, 11, 11, 11, 11,  0])
loss=  tensor(0.0034, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.7586,
3.7586, 3.7586, 3.7586, 3.7586, 3.7586, 3.7586, 3.7586, 3.7586,
        3.7586, 3.7586, 3.7586, 3.7586, 3.7586, 3.7586, 3.7586, 3.7586, 3.2537,
        2.7873, 2.3488, 1.9307, 1.5275, 1.1353, 0.7513, 0.3734]) return=
125571.79187222246
probs of actions:  tensor([0.9984, 0.9992, 0.9983, 0.9992, 0.9992, 0.9989,
0.9992, 0.9995, 0.9994,
        0.9991, 0.9994, 0.9986, 0.9993, 0.9987, 0.9966, 0.9993, 0.9990, 0.9991,
        0.9994, 0.9991, 0.9991, 0.9996, 0.9999, 1.0000, 0.9963],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4923, 0.4922, 0.4922, 0.4921, 0.4921, 0.5042])
finalReturns:  tensor([0.6844, 0.6965, 0.6703, 0.6163, 0.5422, 0.4531, 0.3531,
0.2450, 0.1308])
----------------------------------------
iter  2  stage  15  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11, 11, 11,
        11, 11, 11, 11, 11, 11,  0])
loss=  tensor(0.0043, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.1295,
4.1295, 4.1295, 4.1295, 4.1295, 4.1295, 4.1295, 4.1295, 4.1295,
        4.1295, 4.1295, 4.1295, 4.1295, 4.1295, 4.1295, 4.1295, 3.6244, 3.1578,
        2.7192, 2.3010, 1.8977, 1.5055, 1.1214, 0.7435, 0.3700]) return=
125571.79187222246
probs of actions:  tensor([0.9984, 0.9992, 0.9983, 0.9992, 0.9992, 0.9989,
```

```
0.9992, 0.9995, 0.9994,
        0.9991, 0.9994, 0.9986, 0.9993, 0.9987, 0.9966, 0.9993, 0.9990, 0.9991,
        0.9994, 0.9991, 0.9991, 0.9996, 0.9999, 1.0000, 0.9963],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4923, 0.4922, 0.4922, 0.4921, 0.4921, 0.5042])
finalReturns:  tensor([0.8065, 0.8186, 0.7924, 0.7384, 0.6642, 0.5752, 0.4751,
0.3670, 0.2528,
        0.1341])
----------------------------------------
iter  2  stage  14  ep  11090   adversary:  AdversaryModes.fight_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11, 11, 11,
        11, 11, 11, 11, 11, 11,  0])
loss=  tensor(0.0038, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.4983,
4.4983, 4.4983, 4.4983, 4.4983, 4.4983, 4.4983, 4.4983, 4.4983, 4.4983,
        4.4983, 4.4983, 4.4983, 4.4983, 4.4983, 4.4983, 3.9928, 3.5260, 3.0872,
        2.6688, 2.2654, 1.8732, 1.4891, 1.1111, 0.7376, 0.3675]) return=
125571.79187222246
probs of actions:  tensor([0.9992, 0.9996, 0.9991, 0.9995, 0.9995, 0.9994,
0.9995, 0.9997, 0.9996,
        0.9995, 0.9996, 0.9992, 0.9996, 0.9993, 0.9990, 0.9997, 0.9998, 0.9997,
        0.9997, 0.9996, 0.9995, 0.9998, 1.0000, 1.0000, 0.9916],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4923, 0.4922, 0.4922, 0.4921, 0.4921, 0.5042])
finalReturns:  tensor([0.9311, 0.9432, 0.9170, 0.8630, 0.7888, 0.6997, 0.5997,
0.4915, 0.3774,
        0.2587, 0.1366])
----------------------------------------
iter  2  stage  13  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11, 11, 11,
        11, 11, 11, 11, 11, 11,  0])
loss=  tensor(0.0048, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.8656,
4.8656, 4.8656, 4.8656, 4.8656, 4.8656, 4.8656, 4.8656, 4.8656, 4.8656,
        4.8656, 4.8656, 4.8656, 4.8656, 4.8656, 4.8656, 4.3597, 3.8925, 3.4535, 3.0349,
        2.6314, 2.2390, 1.8549, 1.4768, 1.1033, 0.7332, 0.3657]) return=
125571.79187222246
probs of actions:  tensor([0.9992, 0.9996, 0.9991, 0.9995, 0.9995, 0.9994,
0.9995, 0.9997, 0.9996,
        0.9995, 0.9996, 0.9992, 0.9996, 0.9993, 0.9990, 0.9997, 0.9998, 0.9997,
        0.9997, 0.9996, 0.9995, 0.9998, 1.0000, 1.0000, 0.9916],
       grad_fn=<ExpBackward0>)
```

```
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4923, 0.4922, 0.4922, 0.4921, 0.4921, 0.5042])
finalReturns:  tensor([1.0577, 1.0698, 1.0435, 0.9895, 0.9153, 0.8262, 0.7262,
0.6180, 0.5038,
        0.3851, 0.2631, 0.1385])
---------------------------------------
iter  2  stage  12  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11, 11, 11,
        11, 11, 11, 11, 11, 11, 11])
loss=  tensor(0.6158, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.2321,
5.2321, 5.2321, 5.2321, 5.2321, 5.2321, 5.2321, 5.2321, 5.2321,
        5.2321, 5.2321, 5.2321, 5.2321, 4.7256, 4.2580, 3.8186, 3.3998, 2.9961,
        2.6036, 2.2193, 1.8412, 1.4677, 1.0975, 0.7299, 0.3643]) return=
125450.79187222246
probs of actions:  tensor([0.9992, 0.9996, 0.9991, 0.9995, 0.9995, 0.9994,
0.9995, 0.9997, 0.9996,
        0.9995, 0.9996, 0.9992, 0.9996, 0.9993, 0.9990, 0.9997, 0.9998, 0.9997,
        0.9997, 0.9996, 0.9995, 0.9998, 1.0000, 1.0000, 0.0084],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4923, 0.4922, 0.4922, 0.4921, 0.4921, 0.4921])
finalReturns:  tensor([1.1735, 1.1856, 1.1594, 1.1053, 1.0311, 0.9420, 0.8419,
0.7337, 0.6195,
        0.5008, 0.3788, 0.2542, 0.1278])
---------------------------------------
iter  2  stage  11  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11, 11, 11,
        11, 11, 11, 11, 11, 11,  0])
loss=  tensor(0.0069, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.5984,
5.5984, 5.5984, 5.5984, 5.5984, 5.5984, 5.5984, 5.5984, 5.5984,
        5.5984, 5.5984, 5.5984, 5.0910, 4.6228, 4.1830, 3.7639, 3.3600, 2.9673,
        2.5829, 2.2046, 1.8310, 1.4608, 1.0932, 0.7275, 0.3632]) return=
125571.79187222246
probs of actions:  tensor([0.9992, 0.9996, 0.9991, 0.9995, 0.9995, 0.9994,
0.9995, 0.9997, 0.9996,
        0.9995, 0.9996, 0.9992, 0.9996, 0.9993, 0.9990, 0.9997, 0.9998, 0.9997,
        0.9997, 0.9996, 0.9995, 0.9998, 1.0000, 1.0000, 0.9916],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4923, 0.4922, 0.4922, 0.4921, 0.4921, 0.5042])
```

```
finalReturns:  tensor([1.3147, 1.3268, 1.3005, 1.2465, 1.1722, 1.0830, 0.9830,
0.8748, 0.7605,
        0.6418, 0.5198, 0.3952, 0.2688, 0.1410])
----------------------------------------
iter  2  stage  10  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11, 11, 11,
        11, 11, 11, 11, 11, 11,  0])
loss=  tensor(0.0077, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.9648,
5.9648, 5.9648, 5.9648, 5.9648, 5.9648, 5.9648, 5.9648, 5.9648,
        5.9648, 5.9648, 5.4563, 4.9874, 4.5470, 4.1274, 3.7232, 3.3303, 2.9457,
        2.5674, 2.1937, 1.8234, 1.4557, 1.0900, 0.7257, 0.3624]) return=
125571.79187222246
probs of actions:  tensor([0.9992, 0.9996, 0.9991, 0.9995, 0.9995, 0.9994,
0.9995, 0.9997, 0.9996,
        0.9995, 0.9996, 0.9992, 0.9996, 0.9993, 0.9990, 0.9997, 0.9998, 0.9997,
        0.9997, 0.9996, 0.9995, 0.9998, 1.0000, 1.0000, 0.9916],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4923, 0.4922, 0.4922, 0.4921, 0.4921, 0.5042])
finalReturns:  tensor([1.4446, 1.4567, 1.4304, 1.3763, 1.3020, 1.2128, 1.1127,
1.0045, 0.8902,
        0.7715, 0.6495, 0.5249, 0.3985, 0.2706, 0.1418])
----------------------------------------
iter  2  stage  9  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11, 11, 11,
        11, 11, 11, 11, 11, 11,  0])
loss=  tensor(0.0088, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.3321,
6.3321, 6.3321, 6.3321, 6.3321, 6.3321, 6.3321, 6.3321, 6.3321,
        6.3321, 5.8221, 5.3521, 4.9109, 4.4908, 4.0862, 3.6930, 3.3082, 2.9296,
        2.5558, 2.1855, 1.8177, 1.4519, 1.0875, 0.7243, 0.3618]) return=
125571.79187222246
probs of actions:  tensor([0.9992, 0.9996, 0.9991, 0.9995, 0.9995, 0.9994,
0.9995, 0.9997, 0.9996,
        0.9995, 0.9996, 0.9992, 0.9996, 0.9993, 0.9990, 0.9997, 0.9998, 0.9997,
        0.9997, 0.9996, 0.9995, 0.9998, 1.0000, 1.0000, 0.9916],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4923, 0.4922, 0.4922, 0.4921, 0.4921, 0.5042])
finalReturns:  tensor([1.5753, 1.5874, 1.5610, 1.5068, 1.4325, 1.3433, 1.2431,
1.1348, 1.0206,
        0.9018, 0.7797, 0.6552, 0.5287, 0.4009, 0.2720, 0.1423])
----------------------------------------
```

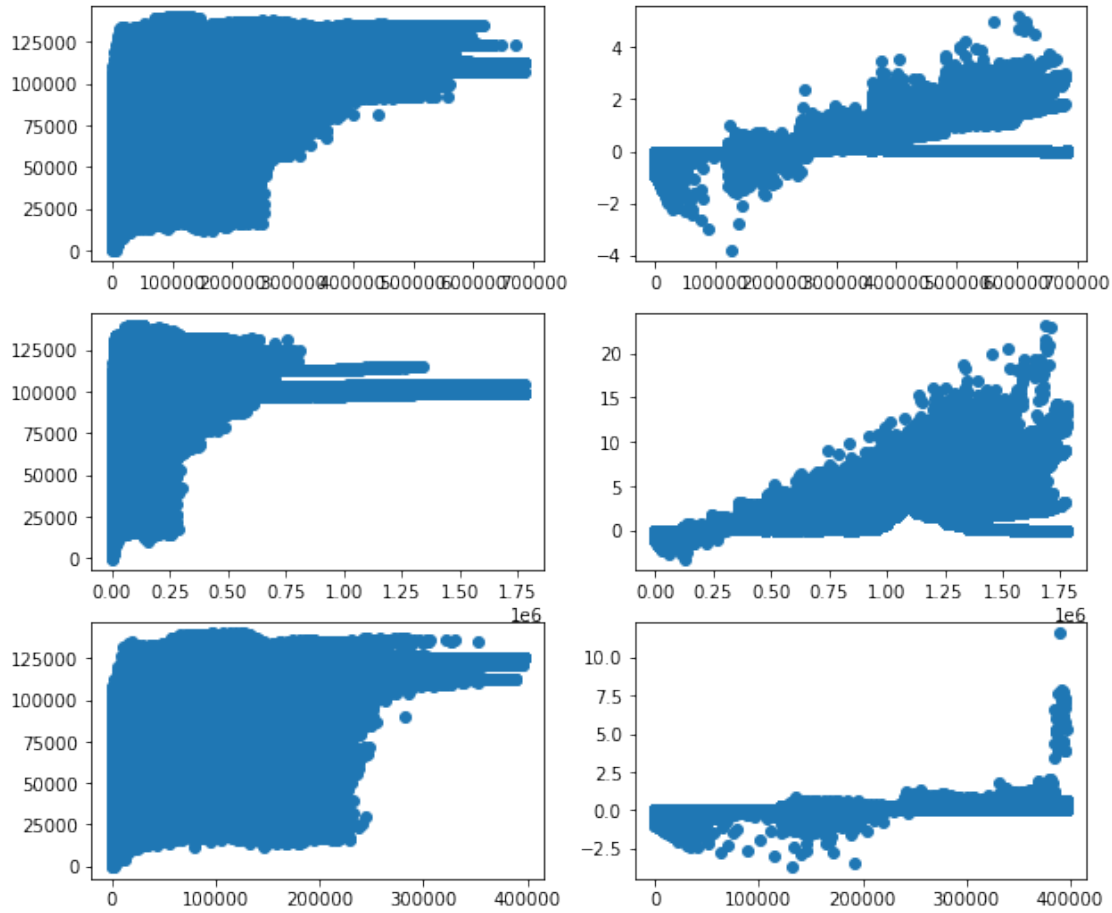
```
iter  2  stage  8  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11, 11, 11,
        11, 11, 11, 11, 11, 11,  0])
loss=  tensor(0.0097, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.7008,
6.7008, 6.7008, 6.7008, 6.7008, 6.7008, 6.7008, 6.7008, 6.7008, 6.7008,
        6.1888, 5.7173, 5.2752, 4.8543, 4.4491, 4.0555, 3.6704, 3.2916, 2.9176,
        2.5471, 2.1793, 1.8134, 1.4490, 1.0857, 0.7232, 0.3614]) return=
125571.79187222246
probs of actions:  tensor([0.9992, 0.9996, 0.9991, 0.9995, 0.9995, 0.9994,
0.9995, 0.9997, 0.9996,
        0.9995, 0.9996, 0.9992, 0.9996, 0.9993, 0.9990, 0.9997, 0.9998, 0.9997,
        0.9997, 0.9996, 0.9995, 0.9998, 1.0000, 1.0000, 0.9916],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4923, 0.4922, 0.4922, 0.4921, 0.4921, 0.5042])
finalReturns:  tensor([1.7065, 1.7186, 1.6921, 1.6379, 1.5635, 1.4742, 1.3740,
1.2657, 1.1514,
        1.0326, 0.9105, 0.7859, 0.6594, 0.5316, 0.4027, 0.2730, 0.1428])
----------------------------------------
iter  2  stage  7  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11, 11, 11,
        11, 11, 11, 11, 11, 11,  0])
loss=  tensor(0.0105, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.0715,
7.0715, 7.0715, 7.0715, 7.0715, 7.0715, 7.0715, 7.0715, 6.5570,
        6.0836, 5.6400, 5.2182, 4.8123, 4.4181, 4.0326, 3.6535, 3.2793, 2.9086,
        2.5407, 2.1747, 1.8102, 1.4469, 1.0844, 0.7225, 0.3611]) return=
125571.79187222246
probs of actions:  tensor([0.9992, 0.9996, 0.9991, 0.9995, 0.9995, 0.9994,
0.9995, 0.9997, 0.9996,
        0.9995, 0.9996, 0.9992, 0.9996, 0.9993, 0.9990, 0.9997, 0.9998, 0.9997,
        0.9997, 0.9996, 0.9995, 0.9998, 1.0000, 1.0000, 0.9915],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4923, 0.4922, 0.4922, 0.4921, 0.4921, 0.5042])
finalReturns:  tensor([1.8382, 1.8503, 1.8238, 1.7694, 1.6949, 1.6055, 1.5052,
1.3969, 1.2825,
        1.1637, 1.0416, 0.9170, 0.7905, 0.6626, 0.5337, 0.4041, 0.2738, 0.1431])
----------------------------------------
iter  2  stage  6  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11, 11, 11,
        11, 11, 11, 11, 11, 11,  0])
```

```
loss=  tensor(0.0117, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.4454,
7.4454, 7.4454, 7.4454, 7.4454, 7.4454, 7.4454, 6.9273, 6.4514,
        6.0060, 5.5828, 5.1759, 4.7809, 4.3949, 4.0155, 3.6409, 3.2701, 2.9019,
        2.5358, 2.1712, 1.8078, 1.4453, 1.0833, 0.7219, 0.3608]) return=
125571.79187222246
probs of actions:  tensor([0.9992, 0.9996, 0.9991, 0.9995, 0.9995, 0.9994,
0.9995, 0.9997, 0.9996,
        0.9995, 0.9996, 0.9992, 0.9996, 0.9993, 0.9990, 0.9997, 0.9998, 0.9997,
        0.9997, 0.9996, 0.9995, 0.9998, 1.0000, 1.0000, 0.9915],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4923, 0.4922, 0.4922, 0.4921, 0.4921, 0.5042])
finalReturns:  tensor([1.9703, 1.9824, 1.9558, 1.9014, 1.8267, 1.7372, 1.6368,
1.5284, 1.4140,
        1.2951, 1.1729, 1.0483, 0.9218, 0.7939, 0.6650, 0.5354, 0.4051, 0.2744,
        0.1434])
----------------------------------------
iter  2  stage  5  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11, 11, 11,
        11, 11, 11, 11, 11, 11,  0])
loss=  tensor(0.0132, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.8234,
7.8234, 7.8234, 7.8234, 7.8234, 7.8234, 7.3006, 6.8213, 6.3735,
        5.9485, 5.5402, 5.1443, 4.7576, 4.3776, 4.0027, 3.6315, 3.2631, 2.8969,
        2.5322, 2.1687, 1.8060, 1.4440, 1.0826, 0.7215, 0.3606]) return=
125571.79187222246
probs of actions:  tensor([0.9992, 0.9996, 0.9991, 0.9995, 0.9995, 0.9994,
0.9995, 0.9997, 0.9996,
        0.9995, 0.9996, 0.9992, 0.9996, 0.9993, 0.9990, 0.9997, 0.9998, 0.9997,
        0.9997, 0.9996, 0.9995, 0.9998, 1.0000, 1.0000, 0.9915],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4923, 0.4922, 0.4922, 0.4921, 0.4921, 0.5042])
finalReturns:  tensor([2.1030, 2.1151, 2.0884, 2.0337, 1.9589, 1.8692, 1.7687,
1.6602, 1.5457,
        1.4268, 1.3045, 1.1799, 1.0533, 0.9254, 0.7965, 0.6668, 0.5366, 0.4059,
        0.2748, 0.1436])
----------------------------------------
iter  2  stage  4  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11, 11, 11,
        11, 11, 11, 11, 11, 11,  0])
loss=  tensor(0.0145, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.2072,
8.2072, 8.2072, 8.2072, 8.2072, 7.6781, 7.1943, 6.7431, 6.3158,
```

```
        5.9058, 5.5086, 5.1208, 4.7401, 4.3647, 3.9931, 3.6244, 3.2580, 2.8931,
        2.5295, 2.1667, 1.8047, 1.4431, 1.0820, 0.7211, 0.3605]) return=
125571.79187222246
probs of actions:  tensor([0.9992, 0.9996, 0.9991, 0.9995, 0.9995, 0.9994,
0.9995, 0.9997, 0.9996,
        0.9995, 0.9996, 0.9992, 0.9996, 0.9993, 0.9990, 0.9997, 0.9998, 0.9997,
        0.9997, 0.9996, 0.9995, 0.9998, 1.0000, 1.0000, 0.9915],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4923, 0.4922, 0.4922, 0.4921, 0.4921, 0.5042])
finalReturns:  tensor([2.2362, 2.2483, 2.2214, 2.1666, 2.0915, 2.0016, 1.9009,
1.7922, 1.6776,
        1.5586, 1.4363, 1.3116, 1.1850, 1.0571, 0.9282, 0.7985, 0.6682, 0.5375,
        0.4064, 0.2752, 0.1437])
----------------------------------------
iter  2  stage  3  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11, 11, 11,
        11, 11, 11, 11, 11, 11,  0])
loss=  tensor(0.0158, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.5988,
8.5988, 8.5988, 8.5988, 8.0612, 7.5713, 7.1158, 6.6852, 6.2729,
        5.8739, 5.4849, 5.1033, 4.7271, 4.3550, 3.9859, 3.6192, 3.2541, 2.8903,
        2.5274, 2.1653, 1.8037, 1.4425, 1.0816, 0.7209, 0.3604]) return=
125571.79187222246
probs of actions:  tensor([0.9992, 0.9996, 0.9991, 0.9995, 0.9995, 0.9994,
0.9995, 0.9997, 0.9996,
        0.9995, 0.9996, 0.9992, 0.9996, 0.9993, 0.9990, 0.9997, 0.9998, 0.9997,
        0.9997, 0.9996, 0.9995, 0.9998, 1.0000, 1.0000, 0.9915],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4923, 0.4922, 0.4922, 0.4921, 0.4921, 0.5042])
finalReturns:  tensor([2.3701, 2.3822, 2.3550, 2.2999, 2.2245, 2.1343, 2.0334,
1.9246, 1.8098,
        1.6907, 1.5683, 1.4435, 1.3169, 1.1889, 1.0599, 0.9302, 0.7999, 0.6692,
        0.5381, 0.4069, 0.2754, 0.1438])
----------------------------------------
iter  2  stage  2  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11, 11, 11,
        11, 11, 11, 11, 11, 11,  0])
loss=  tensor(0.0183, grad_fn=<NegBackward0>)   ,  base rewards= tensor([9.0010,
9.0010, 9.0010, 8.4520, 7.9540, 7.4926, 7.0577, 6.6422, 6.2409,
        5.8502, 5.4673, 5.0901, 4.7173, 4.3477, 3.9806, 3.6152, 3.2512, 2.8881,
        2.5259, 2.1642, 1.8029, 1.4420, 1.0812, 0.7207, 0.3603]) return=
```

```
125571.79187222246
probs of actions:  tensor([0.9992, 0.9996, 0.9991, 0.9995, 0.9995, 0.9994,
0.9995, 0.9997, 0.9996,
        0.9995, 0.9996, 0.9992, 0.9996, 0.9993, 0.9990, 0.9997, 0.9998, 0.9997,
        0.9997, 0.9996, 0.9995, 0.9998, 1.0000, 1.0000, 0.9915],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4923, 0.4922, 0.4922, 0.4921, 0.4921, 0.5042])
finalReturns:  tensor([2.5048, 2.5169, 2.4894, 2.4338, 2.3580, 2.2675, 2.1663,
2.0572, 1.9422,
        1.8229, 1.7004, 1.5755, 1.4489, 1.3208, 1.1918, 1.0621, 0.9317, 0.8010,
        0.6699, 0.5387, 0.4072, 0.2756, 0.1439])
----------------------------------------
iter  2  stage  1  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11, 11, 11,
        11, 11, 11, 11, 11, 11,  0])
loss=  tensor(0.0198, grad_fn=<NegBackward0>)   ,  base rewards= tensor([9.4176,
9.4176, 8.8532, 8.3442, 7.8749, 7.4342, 7.0145, 6.6101, 6.2171,
        5.8325, 5.4541, 5.0803, 4.7100, 4.3423, 3.9765, 3.6122, 3.2490, 2.8866,
        2.5247, 2.1634, 1.8023, 1.4416, 1.0810, 0.7206, 0.3602]) return=
125571.79187222246
probs of actions:  tensor([0.9992, 0.9996, 0.9991, 0.9995, 0.9995, 0.9994,
0.9995, 0.9997, 0.9996,
        0.9995, 0.9996, 0.9992, 0.9996, 0.9993, 0.9990, 0.9997, 0.9998, 0.9997,
        0.9997, 0.9996, 0.9995, 0.9998, 1.0000, 1.0000, 0.9915],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4923, 0.4922, 0.4922, 0.4921, 0.4921, 0.5042])
finalReturns:  tensor([2.6405, 2.6526, 2.6247, 2.5685, 2.4922, 2.4012, 2.2995,
2.1901, 2.0749,
        1.9554, 1.8328, 1.7077, 1.5810, 1.4529, 1.3238, 1.1940, 1.0637, 0.9329,
        0.8018, 0.6705, 0.5390, 0.4074, 0.2757, 0.1439])
----------------------------------------
iter  2  stage  0  ep  0   adversary:  AdversaryModes.fight_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11, 11, 11,
        11, 11, 11, 11, 11, 11,  0])
loss=  tensor(0.0224, grad_fn=<NegBackward0>)   ,  base rewards= tensor([9.7796,
9.2684, 8.7446, 8.2646, 7.8162, 7.3909, 6.9823, 6.5862, 6.1993,
        5.8192, 5.4442, 5.0730, 4.7046, 4.3383, 3.9735, 3.6100, 3.2473, 2.8854,
        2.5239, 2.1628, 1.8019, 1.4413, 1.0808, 0.7205, 0.3602]) return=
125571.79187222246
probs of actions:  tensor([0.9992, 0.9996, 0.9991, 0.9995, 0.9996, 0.9994,
```

0.9995, 0.9997, 0.9996,
        0.9995, 0.9996, 0.9992, 0.9996, 0.9993, 0.9990, 0.9997, 0.9998, 0.9997,
        0.9997, 0.9996, 0.9995, 0.9998, 1.0000, 1.0000, 0.9915],
       grad_fn=<ExpBackward0>)
rewards: tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4923, 0.4922, 0.4922, 0.4921, 0.4921, 0.5042])
finalReturns: tensor([2.7775, 2.7896, 2.7612, 2.7043, 2.6272, 2.5355, 2.4333,
2.3235, 2.2079,
        2.0881, 1.9653, 1.8401, 1.7132, 1.5850, 1.4559, 1.3260, 1.1956, 1.0649,
        0.9338, 0.8024, 0.6709, 0.5393, 0.4076, 0.2758, 0.1440])
0,[1e-05,1][1, 10000, 1, 1],1682983428 saved
[396616, 'tensor([0., 0., 0., 0., 0., 1., 0., 0., 0., 0., 0., 0.])',
125571.79187222246, 95728.31346963784, 0.022367581725120544, 1e-05, 1, 0,
'tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11,\n   11, 11, 11, 11, 11, 11, 0])', '[1.   1.   1.   1.   1.   1.   1.
1.   1.   1.   1.   1.   1.   1.   1.\n 1.   1.   1.   1.   1.   1.   1.   1.
1.   0.99]', '0,[1e-05,1][1, 10000, 1, 1],1682983428', 25, 50,
156312.25008636713, 172508.0785299521, 67091.50476852678, 135143.75466666667,
132277.408, 125571.79187222247, 125571.79187222247, 128010.77984270274,
128010.77984270274, 78103.35371185312, 125571.79187222247, 128010.77984270274]




```
policy reset
----------------------------------------
iter  0  stage  24  ep  99999   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0,
        0])
loss=  tensor(-0., grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.5624,
0.5624, 0.5624, 0.5624, 0.5624, 0.5624, 0.5624, 0.5624, 0.5624,
        0.5624, 0.5624, 0.5624, 0.5624, 0.5624, 0.5624, 0.5624, 0.5624, 0.5624,
        0.5624, 0.5624, 0.5624, 0.5624, 0.5624, 0.5624, 0.5624]) return=
138554.5803254051
probs of actions:  tensor([0.8580, 0.8369, 0.8854, 0.8691, 0.8611, 0.8791,
0.8596, 0.8337, 0.8636,
        0.8863, 0.8837, 0.8661, 0.8775, 0.8598, 0.8728, 0.8699, 0.8710, 0.8759,
        0.8779, 0.8717, 0.8976, 0.8724, 0.8831, 0.8544, 0.9742],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5112, 0.5238, 0.5334, 0.5406, 0.5460, 0.5501, 0.5532,
0.5555, 0.5573,
        0.5586, 0.5595, 0.5603, 0.5608, 0.5613, 0.5616, 0.5618, 0.5620, 0.5621,
```

```
        0.5622, 0.5623, 0.5623, 0.5624, 0.5624, 0.5624, 0.5624])
finalReturns:  tensor([0.])
----------------------------------------
iter  0  stage  23  ep  99999   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([13, 12, 11,  0,  0,  9, 11, 12, 17,  8, 13,  0,  9,  9, 10,
0, 10, 10,
        20, 10,  0,  0, 16, 12,  0])
loss=  tensor(0.0255, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.8069,
0.8069, 0.8069, 0.8069, 0.8069, 0.8069, 0.8069, 0.8069, 0.8069,
        0.8069, 0.8069, 0.8069, 0.8069, 0.8069, 0.8069, 0.8069, 0.8069, 0.8069,
        0.8069, 0.8069, 0.8069, 0.8069, 0.8069, 0.8069, 0.3898]) return=
108454.93368961285
probs of actions:  tensor([0.1475, 0.0568, 0.0506, 0.2173, 0.2489, 0.1672,
0.0495, 0.0950, 0.0052,
        0.0350, 0.1080, 0.1715, 0.1418, 0.1597, 0.0460, 0.3593, 0.0511, 0.0653,
        0.0013, 0.0516, 0.3377, 0.2581, 0.0042, 0.3484, 0.9947],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4943, 0.5575, 0.5313, 0.5190, 0.4628, 0.4147, 0.4106,
0.4148, 0.4085,
        0.4540, 0.4305, 0.4545, 0.4087, 0.4101, 0.4093, 0.4233, 0.3844, 0.3945,
        0.3722, 0.4409, 0.4471, 0.4115, 0.3602, 0.4027, 0.4282])
finalReturns:  tensor([0.0240, 0.0384])
----------------------------------------
iter  0  stage  22  ep  99999   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([ 7, 13, 13, 13, 13, 13, 13, 13, 12,  0, 13, 13, 13, 17, 13,
13, 13, 12,
        13, 13, 13, 13, 13, 12,  0])
loss=  tensor(0.0636, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.1083,
1.1083, 1.1083, 1.1083, 1.1083, 1.1083, 1.1083, 1.1083, 1.1083,
        1.1083, 1.1083, 1.1083, 1.1083, 1.1083, 1.1083, 1.1083, 1.1083, 1.1083,
        1.1083, 1.1083, 1.1083, 1.1083, 1.1083, 0.7025, 0.3368]) return=
106345.43462592686
probs of actions:  tensor([0.0019, 0.8371, 0.8408, 0.8583, 0.8412, 0.8676,
0.8239, 0.7916, 0.1241,
        0.0167, 0.8616, 0.8110, 0.8468, 0.0015, 0.8177, 0.8115, 0.8583, 0.1668,
        0.7993, 0.8224, 0.8249, 0.8218, 0.9099, 0.5416, 0.9995],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5063, 0.5326, 0.5852, 0.5327, 0.4949, 0.4674, 0.4473,
0.4325, 0.4240,
        0.4270, 0.3638, 0.3701, 0.3749, 0.3665, 0.3940, 0.3929, 0.3920, 0.3939,
        0.3877, 0.3882, 0.3885, 0.3888, 0.3889, 0.3916, 0.4029])
finalReturns:  tensor([0.0751, 0.0920, 0.0661])
----------------------------------------
iter  0  stage  21  ep  99999   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([13, 13, 13, 13, 13, 13, 14, 13, 13, 13, 13, 13, 13, 13, 13,
13, 13, 13,
        12, 13, 13, 14, 13, 12,  0])
loss=  tensor(1.2667, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.6831,
```

```
1.6831, 1.6831, 1.6831, 1.6831, 1.6831, 1.6831, 1.6831, 1.6831,
        1.6831, 1.6831, 1.6831, 1.6831, 1.6831, 1.6831, 1.6831, 1.6831, 1.6831,
        1.6831, 1.6831, 1.6831, 1.6831, 1.2086, 0.7775, 0.3774]) return=
119140.59757259364
probs of actions:  tensor([0.9729, 0.9534, 0.9539, 0.9617, 0.9595, 0.9673,
0.0031, 0.9340, 0.9457,
        0.9719, 0.9712, 0.9486, 0.9622, 0.9698, 0.9563, 0.9580, 0.9652, 0.9294,
        0.0450, 0.9563, 0.9583, 0.0017, 0.9820, 0.3256, 0.9998],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4943, 0.5550, 0.5302, 0.5120, 0.4986, 0.4886, 0.4784,
0.4791, 0.4741,
        0.4704, 0.4676, 0.4655, 0.4639, 0.4627, 0.4618, 0.4612, 0.4607, 0.4603,
        0.4625, 0.4564, 0.4571, 0.4549, 0.4615, 0.4634, 0.4739])
finalReturns:  tensor([0.1705, 0.1901, 0.1599, 0.0965])
----------------------------------------
iter  0  stage  20  ep  99999   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 12, 13, 13, 13,
13, 13, 13,
        13, 13, 13, 13, 13, 13,  0])
loss=  tensor(0.0540, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.0488,
2.0488, 2.0488, 2.0488, 2.0488, 2.0488, 2.0488, 2.0488, 2.0488,
        2.0488, 2.0488, 2.0488, 2.0488, 2.0488, 2.0488, 2.0488, 2.0488, 2.0488,
        2.0488, 2.0488, 2.0488, 1.5726, 1.1403, 0.7394, 0.3614]) return=
118964.00360478487
probs of actions:  tensor([0.9881, 0.9788, 0.9789, 0.9827, 0.9821, 0.9856,
0.9772, 0.9685, 0.9748,
        0.9878, 0.9878, 0.0196, 0.9830, 0.9867, 0.9804, 0.9813, 0.9843, 0.9665,
        0.9737, 0.9803, 0.9839, 0.9912, 0.9935, 0.7924, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4943, 0.5550, 0.5302, 0.5120, 0.4986, 0.4886, 0.4811,
0.4756, 0.4715,
        0.4684, 0.4661, 0.4669, 0.4596, 0.4595, 0.4594, 0.4594, 0.4593, 0.4593,
        0.4593, 0.4593, 0.4592, 0.4592, 0.4592, 0.4592, 0.4761])
finalReturns:  tensor([0.2643, 0.2812, 0.2543, 0.1959, 0.1147])
----------------------------------------
iter  0  stage  19  ep  58190   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13,
13, 13, 13,
        13, 13, 13, 13, 13, 13,  0])
loss=  tensor(0.0039, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.3993,
2.3993, 2.3993, 2.3993, 2.3993, 2.3993, 2.3993, 2.3993, 2.3993,
        2.3993, 2.3993, 2.3993, 2.3993, 2.3993, 2.3993, 2.3993, 2.3993, 2.3993,
        2.3993, 2.3993, 1.9227, 1.4901, 1.0889, 0.7107, 0.3492]) return=
119073.90192566637
probs of actions:  tensor([0.9991, 0.9983, 0.9984, 0.9987, 0.9986, 0.9990,
0.9981, 0.9974, 0.9981,
        0.9992, 0.9991, 0.9981, 0.9987, 0.9989, 0.9984, 0.9985, 0.9988, 0.9975,
        0.9980, 0.9990, 0.9994, 0.9997, 0.9998, 0.9864, 1.0000],
```

```
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4943, 0.5550, 0.5302, 0.5120, 0.4986, 0.4886, 0.4811,
0.4756, 0.4715,
        0.4684, 0.4661, 0.4644, 0.4631, 0.4621, 0.4614, 0.4608, 0.4604, 0.4601,
        0.4599, 0.4597, 0.4596, 0.4595, 0.4594, 0.4594, 0.4762])
finalReturns:  tensor([0.3745, 0.3914, 0.3644, 0.3061, 0.2249, 0.1270])
----------------------------------------
iter  0  stage  18  ep  1091   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13,
13, 13, 13,
        13, 13, 13, 13, 13, 13,  0])
loss=  tensor(0.0049, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.7400,
2.7400, 2.7400, 2.7400, 2.7400, 2.7400, 2.7400, 2.7400, 2.7400,
        2.7400, 2.7400, 2.7400, 2.7400, 2.7400, 2.7400, 2.7400, 2.7400, 2.7400,
        2.7400, 2.2632, 1.8304, 1.4292, 1.0509, 0.6894, 0.3402]) return=
119073.90192566637
probs of actions:  tensor([0.9991, 0.9983, 0.9984, 0.9987, 0.9986, 0.9990,
0.9982, 0.9974, 0.9981,
        0.9992, 0.9992, 0.9981, 0.9987, 0.9990, 0.9984, 0.9985, 0.9988, 0.9975,
        0.9990, 0.9990, 0.9994, 0.9998, 0.9998, 0.9863, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4943, 0.5550, 0.5302, 0.5120, 0.4986, 0.4886, 0.4811,
0.4756, 0.4715,
        0.4684, 0.4661, 0.4644, 0.4631, 0.4621, 0.4614, 0.4608, 0.4604, 0.4601,
        0.4599, 0.4597, 0.4596, 0.4595, 0.4594, 0.4594, 0.4762])
finalReturns:  tensor([0.4937, 0.5106, 0.4836, 0.4253, 0.3441, 0.2462, 0.1361])
----------------------------------------
iter  0  stage  17  ep  11404   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([11, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13,
13, 13, 13,
        13, 13, 13, 13, 13, 13,  0])
loss=  tensor(0.0041, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.4808,
3.4808, 3.4808, 3.4808, 3.4808, 3.4808, 3.4808, 3.4808, 3.4808,
        3.4808, 3.4808, 3.4808, 3.4808, 3.4808, 3.4808, 3.4808, 3.4808, 3.4808,
        2.9476, 2.4609, 2.0076, 1.5787, 1.1675, 0.7695, 0.3812]) return=
130250.32620960443
probs of actions:  tensor([1.3511e-04, 9.9879e-01, 9.9887e-01, 9.9912e-01,
9.9907e-01, 9.9933e-01,
        9.9877e-01, 9.9826e-01, 9.9875e-01, 9.9946e-01, 9.9946e-01, 9.9876e-01,
        9.9916e-01, 9.9934e-01, 9.9897e-01, 9.9902e-01, 9.9923e-01, 9.9906e-01,
        9.9938e-01, 9.9941e-01, 9.9971e-01, 9.9991e-01, 9.9989e-01, 9.9131e-01,
        1.0000e+00], grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5475, 0.5395, 0.5336, 0.5292, 0.5259, 0.5234,
0.5215, 0.5201,
        0.5191, 0.5183, 0.5177, 0.5173, 0.5170, 0.5167, 0.5166, 0.5164, 0.5163,
        0.5162, 0.5162, 0.5161, 0.5161, 0.5161, 0.5161, 0.5329])
finalReturns:  tensor([0.6652, 0.6821, 0.6526, 0.5897, 0.5025, 0.3975, 0.2795,
0.1518])
```

```
---------------------------------------
iter  0  stage  16  ep  0   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13,
13, 13, 13,
        13, 13, 13, 13, 13, 13,  0])
loss=  tensor(0.0054, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.4037,
3.4037, 3.4037, 3.4037, 3.4037, 3.4037, 3.4037, 3.4037, 3.4037,
        3.4037, 3.4037, 3.4037, 3.4037, 3.4037, 3.4037, 3.4037, 3.4037, 2.9264,
        2.4932, 2.0917, 1.7132, 1.3515, 1.0022, 0.6619, 0.3284]) return=
119073.90192566637
probs of actions:  tensor([0.9993, 0.9988, 0.9989, 0.9991, 0.9990, 0.9993,
0.9987, 0.9981, 0.9986,
        0.9994, 0.9994, 0.9986, 0.9991, 0.9993, 0.9988, 0.9989, 0.9991, 0.9989,
        0.9993, 0.9993, 0.9997, 0.9999, 0.9999, 0.9904, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4943, 0.5550, 0.5302, 0.5120, 0.4986, 0.4886, 0.4811,
0.4756, 0.4715,
        0.4684, 0.4661, 0.4644, 0.4631, 0.4621, 0.4614, 0.4608, 0.4604, 0.4601,
        0.4599, 0.4597, 0.4596, 0.4595, 0.4594, 0.4594, 0.4762])
finalReturns:  tensor([0.7505, 0.7674, 0.7405, 0.6821, 0.6009, 0.5030, 0.3928,
0.2737, 0.1478])
---------------------------------------
iter  0  stage  15  ep  38   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13,
13, 13, 13,
        13, 13, 13, 13, 13, 13,  0])
loss=  tensor(0.0065, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.7298,
3.7298, 3.7298, 3.7298, 3.7298, 3.7298, 3.7298, 3.7298, 3.7298,
        3.7298, 3.7298, 3.7298, 3.7298, 3.7298, 3.7298, 3.7298, 3.2521, 2.8186,
        2.4169, 2.0383, 1.6765, 1.3270, 0.9867, 0.6532, 0.3247]) return=
119073.90192566637
probs of actions:  tensor([0.9994, 0.9988, 0.9989, 0.9991, 0.9990, 0.9993,
0.9987, 0.9981, 0.9986,
        0.9994, 0.9994, 0.9986, 0.9991, 0.9993, 0.9989, 0.9990, 0.9992, 0.9990,
        0.9993, 0.9993, 0.9997, 0.9999, 0.9999, 0.9905, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4943, 0.5550, 0.5302, 0.5120, 0.4986, 0.4886, 0.4811,
0.4756, 0.4715,
        0.4684, 0.4661, 0.4644, 0.4631, 0.4621, 0.4614, 0.4608, 0.4604, 0.4601,
        0.4599, 0.4597, 0.4596, 0.4595, 0.4594, 0.4594, 0.4762])
finalReturns:  tensor([0.8853, 0.9022, 0.8752, 0.8168, 0.7355, 0.6376, 0.5275,
0.4083, 0.2824,
        0.1515])
---------------------------------------
iter  0  stage  14  ep  64   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13,
13, 13, 13,
        13, 13, 13, 13, 13, 13,  0])
```

```
loss=  tensor(0.0078, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.0537,
4.0537, 4.0537, 4.0537, 4.0537, 4.0537, 4.0537, 4.0537, 4.0537,
        4.0537, 4.0537, 4.0537, 4.0537, 4.0537, 4.0537, 3.5754, 3.1415, 2.7396,
        2.3607, 1.9987, 1.6492, 1.3088, 0.9752, 0.6467, 0.3220]) return=
119073.90192566637
probs of actions:  tensor([0.9994, 0.9988, 0.9989, 0.9991, 0.9990, 0.9993,
0.9987, 0.9981, 0.9987,
        0.9994, 0.9994, 0.9987, 0.9991, 0.9993, 0.9990, 0.9991, 0.9993, 0.9990,
        0.9993, 0.9994, 0.9997, 0.9999, 0.9999, 0.9906, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4943, 0.5550, 0.5302, 0.5120, 0.4986, 0.4886, 0.4811,
0.4756, 0.4715,
        0.4684, 0.4661, 0.4644, 0.4631, 0.4621, 0.4614, 0.4608, 0.4604, 0.4601,
        0.4599, 0.4597, 0.4596, 0.4595, 0.4594, 0.4594, 0.4762])
finalReturns:  tensor([1.0228, 1.0397, 1.0127, 0.9542, 0.8730, 0.7751, 0.6649,
0.5457, 0.4198,
        0.2889, 0.1543])
---------------------------------------
iter  0  stage  13  ep  0   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13,
13, 13, 13,
        13, 13, 13, 13, 13, 13,  0])
loss=  tensor(0.0091, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.3761,
4.3761, 4.3761, 4.3761, 4.3761, 4.3761, 4.3761, 4.3761, 4.3761,
        4.3761, 4.3761, 4.3761, 4.3761, 4.3761, 3.8971, 3.4627, 3.0604, 2.6813,
        2.3191, 1.9694, 1.6289, 1.2953, 0.9667, 0.6419, 0.3199]) return=
119073.90192566637
probs of actions:  tensor([0.9994, 0.9988, 0.9989, 0.9991, 0.9990, 0.9993,
0.9987, 0.9982, 0.9987,
        0.9994, 0.9994, 0.9987, 0.9991, 0.9993, 0.9990, 0.9991, 0.9993, 0.9990,
        0.9993, 0.9994, 0.9997, 0.9999, 0.9999, 0.9906, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4943, 0.5550, 0.5302, 0.5120, 0.4986, 0.4886, 0.4811,
0.4756, 0.4715,
        0.4684, 0.4661, 0.4644, 0.4631, 0.4621, 0.4614, 0.4608, 0.4604, 0.4601,
        0.4599, 0.4597, 0.4596, 0.4595, 0.4594, 0.4594, 0.4762])
finalReturns:  tensor([1.1624, 1.1793, 1.1523, 1.0938, 1.0125, 0.9146, 0.8044,
0.6852, 0.5592,
        0.4283, 0.2937, 0.1563])
---------------------------------------
iter  0  stage  12  ep  0   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13,
13, 13, 13,
        13, 13, 13, 13, 13, 13,  0])
loss=  tensor(0.0108, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.6979,
4.6979, 4.6979, 4.6979, 4.6979, 4.6979, 4.6979, 4.6979, 4.6979,
        4.6979, 4.6979, 4.6979, 4.6979, 4.2179, 3.7828, 3.3800, 3.0005, 2.6381,
        2.2882, 1.9476, 1.6138, 1.2851, 0.9603, 0.6382, 0.3183]) return=
```

```
119073.90192566637
probs of actions:  tensor([0.9994, 0.9988, 0.9989, 0.9991, 0.9990, 0.9993,
0.9987, 0.9982, 0.9987,
        0.9994, 0.9994, 0.9987, 0.9991, 0.9993, 0.9990, 0.9991, 0.9993, 0.9990,
        0.9993, 0.9994, 0.9997, 0.9999, 0.9999, 0.9906, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4943, 0.5550, 0.5302, 0.5120, 0.4986, 0.4886, 0.4811,
0.4756, 0.4715,
        0.4684, 0.4661, 0.4644, 0.4631, 0.4621, 0.4614, 0.4608, 0.4604, 0.4601,
        0.4599, 0.4597, 0.4596, 0.4595, 0.4594, 0.4594, 0.4762])
finalReturns:  tensor([1.3037, 1.3206, 1.2936, 1.2350, 1.1537, 1.0557, 0.9455,
0.8262, 0.7003,
        0.5694, 0.4347, 0.2973, 0.1579])
----------------------------------------
iter  0  stage  11  ep  197   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13,
13, 13, 13,
        13, 13, 13, 13, 13, 13,  0])
loss=  tensor(0.0113, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.0197,
5.0197, 5.0197, 5.0197, 5.0197, 5.0197, 5.0197, 5.0197, 5.0197,
        5.0197, 5.0197, 5.0197, 4.5384, 4.1024, 3.6989, 3.3190, 2.9562, 2.6060,
        2.2652, 1.9313, 1.6025, 1.2776, 0.9555, 0.6355, 0.3172]) return=
119073.90192566637
probs of actions:  tensor([0.9994, 0.9989, 0.9989, 0.9992, 0.9991, 0.9993,
0.9988, 0.9982, 0.9987,
        0.9994, 0.9994, 0.9990, 0.9993, 0.9995, 0.9992, 0.9992, 0.9994, 0.9990,
        0.9994, 0.9994, 0.9997, 0.9999, 0.9999, 0.9910, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4943, 0.5550, 0.5302, 0.5120, 0.4986, 0.4886, 0.4811,
0.4756, 0.4715,
        0.4684, 0.4661, 0.4644, 0.4631, 0.4621, 0.4614, 0.4608, 0.4604, 0.4601,
        0.4599, 0.4597, 0.4596, 0.4595, 0.4594, 0.4594, 0.4762])
finalReturns:  tensor([1.4463, 1.4632, 1.4361, 1.3775, 1.2961, 1.1980, 1.0878,
0.9685, 0.8425,
        0.7116, 0.5769, 0.4395, 0.3001, 0.1591])
----------------------------------------
iter  0  stage  10  ep  0   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13,
13, 13, 13,
        13, 13, 13, 13, 13, 13,  0])
loss=  tensor(0.0126, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.3422,
5.3422, 5.3422, 5.3422, 5.3422, 5.3422, 5.3422, 5.3422, 5.3422,
        5.3422, 5.3422, 4.8592, 4.4220, 4.0176, 3.6370, 3.2738, 2.9233, 2.5822,
        2.2481, 1.9191, 1.5941, 1.2719, 0.9519, 0.6335, 0.3163]) return=
119073.90192566637
probs of actions:  tensor([0.9994, 0.9989, 0.9989, 0.9992, 0.9991, 0.9993,
0.9988, 0.9982, 0.9987,
        0.9994, 0.9994, 0.9990, 0.9993, 0.9995, 0.9992, 0.9992, 0.9994, 0.9990,
```

```
        0.9994, 0.9994, 0.9997, 0.9999, 0.9999, 0.9910, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4943, 0.5550, 0.5302, 0.5120, 0.4986, 0.4886, 0.4811,
0.4756, 0.4715,
        0.4684, 0.4661, 0.4644, 0.4631, 0.4621, 0.4614, 0.4608, 0.4604, 0.4601,
        0.4599, 0.4597, 0.4596, 0.4595, 0.4594, 0.4594, 0.4762])
finalReturns:  tensor([1.5898, 1.6067, 1.5796, 1.5209, 1.4394, 1.3413, 1.2310,
1.1116, 0.9856,
        0.8546, 0.7200, 0.5826, 0.4431, 0.3021, 0.1599])
---------------------------------------
iter  0  stage  9  ep  0   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13,
13, 13, 13,
        13, 13, 13, 13, 13, 13,  0])
loss=  tensor(0.0140, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.6662,
5.6662, 5.6662, 5.6662, 5.6662, 5.6662, 5.6662, 5.6662, 5.6662, 5.6662,
        5.6662, 5.1809, 4.7421, 4.3365, 3.9550, 3.5911, 3.2402, 2.8987, 2.5644,
        2.2353, 1.9101, 1.5878, 1.2677, 0.9492, 0.6320, 0.3156]) return=
119073.90192566637
probs of actions:  tensor([0.9994, 0.9989, 0.9989, 0.9992, 0.9991, 0.9993,
0.9988, 0.9982, 0.9987,
        0.9994, 0.9994, 0.9990, 0.9993, 0.9995, 0.9992, 0.9992, 0.9994, 0.9990,
        0.9994, 0.9994, 0.9997, 0.9999, 0.9999, 0.9910, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4943, 0.5550, 0.5302, 0.5120, 0.4986, 0.4886, 0.4811,
0.4756, 0.4715,
        0.4684, 0.4661, 0.4644, 0.4631, 0.4621, 0.4614, 0.4608, 0.4604, 0.4601,
        0.4599, 0.4597, 0.4596, 0.4595, 0.4594, 0.4594, 0.4762])
finalReturns:  tensor([1.7342, 1.7511, 1.7239, 1.6651, 1.5835, 1.4853, 1.3749,
1.2555, 1.1294,
        0.9984, 0.8637, 0.7263, 0.5868, 0.4458, 0.3036, 0.1606])
---------------------------------------
iter  0  stage  8  ep  109   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13,
13, 13, 13,
        13, 13, 13, 13, 13, 13,  0])
loss=  tensor(0.0145, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.9926,
5.9926, 5.9926, 5.9926, 5.9926, 5.9926, 5.9926, 5.9926, 5.9926, 5.9926,
        5.5042, 5.0631, 4.6560, 4.2733, 3.9086, 3.5570, 3.2151, 2.8804, 2.5511,
        2.2257, 1.9033, 1.5831, 1.2645, 0.9472, 0.6308, 0.3152]) return=
119073.90192566637
probs of actions:  tensor([0.9994, 0.9989, 0.9990, 0.9992, 0.9992, 0.9994,
0.9988, 0.9984, 0.9990,
        0.9996, 0.9996, 0.9991, 0.9994, 0.9995, 0.9993, 0.9992, 0.9995, 0.9991,
        0.9994, 0.9994, 0.9997, 0.9999, 0.9999, 0.9917, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4943, 0.5550, 0.5302, 0.5120, 0.4986, 0.4886, 0.4811,
0.4756, 0.4715,
```

```
        0.4684, 0.4661, 0.4644, 0.4631, 0.4621, 0.4614, 0.4608, 0.4604, 0.4601,
        0.4599, 0.4597, 0.4596, 0.4595, 0.4594, 0.4594, 0.4762])
finalReturns:  tensor([1.8793, 1.8962, 1.8689, 1.8100, 1.7283, 1.6299, 1.5194,
1.3999, 1.2738,
        1.1427, 1.0080, 0.8705, 0.7310, 0.5900, 0.4478, 0.3047, 0.1611])
----------------------------------------
iter  0  stage  7  ep  3450   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13,
13, 13, 13,
        13, 13, 13, 13, 13, 13,  0])
loss=  tensor(0.0113, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.3224,
6.3224, 6.3224, 6.3224, 6.3224, 6.3224, 6.3224, 6.3224, 5.8299,
        5.3859, 4.9766, 4.5924, 4.2265, 3.8741, 3.5316, 3.1965, 2.8668, 2.5411,
        2.2185, 1.8982, 1.5795, 1.2622, 0.9457, 0.6300, 0.3148]) return=
119073.90192566637
probs of actions:  tensor([0.9995, 0.9991, 0.9992, 0.9993, 0.9993, 0.9995,
0.9990, 0.9990, 0.9994,
        0.9998, 0.9996, 0.9996, 0.9998, 0.9999, 0.9994, 0.9993, 0.9998, 0.9994,
        0.9995, 0.9996, 0.9999, 1.0000, 0.9999, 0.9931, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4943, 0.5550, 0.5302, 0.5120, 0.4986, 0.4886, 0.4811,
0.4756, 0.4715,
        0.4684, 0.4661, 0.4644, 0.4631, 0.4621, 0.4614, 0.4608, 0.4604, 0.4601,
        0.4599, 0.4597, 0.4596, 0.4595, 0.4594, 0.4594, 0.4762])
finalReturns:  tensor([2.0251, 2.0420, 2.0146, 1.9555, 1.8736, 1.7750, 1.6644,
1.5448, 1.4186,
        1.2874, 1.1527, 1.0152, 0.8756, 0.7345, 0.5923, 0.4493, 0.3056, 0.1614])
----------------------------------------
iter  0  stage  6  ep  0   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13,
13, 13, 13,
        13, 13, 13, 13, 13, 13,  0])
loss=  tensor(0.0138, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.6570,
6.6570, 6.6570, 6.6570, 6.6570, 6.6570, 6.6570, 6.1590, 5.7110,
        5.2989, 4.9127, 4.5453, 4.1918, 3.8484, 3.5127, 3.1825, 2.8566, 2.5337,
        2.2132, 1.8944, 1.5769, 1.2604, 0.9446, 0.6293, 0.3145]) return=
119073.90192566637
probs of actions:  tensor([0.9995, 0.9991, 0.9992, 0.9993, 0.9993, 0.9995,
0.9990, 0.9990, 0.9994,
        0.9998, 0.9996, 0.9996, 0.9998, 0.9999, 0.9994, 0.9993, 0.9998, 0.9994,
        0.9995, 0.9996, 0.9999, 1.0000, 0.9999, 0.9931, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4943, 0.5550, 0.5302, 0.5120, 0.4986, 0.4886, 0.4811,
0.4756, 0.4715,
        0.4684, 0.4661, 0.4644, 0.4631, 0.4621, 0.4614, 0.4608, 0.4604, 0.4601,
        0.4599, 0.4597, 0.4596, 0.4595, 0.4594, 0.4594, 0.4762])
finalReturns:  tensor([2.1717, 2.1886, 2.1609, 2.1016, 2.0194, 1.9207, 1.8098,
1.6901, 1.5637,
```

```
        1.4325, 1.2977, 1.1601, 1.0205, 0.8794, 0.7372, 0.5941, 0.4504, 0.3062,
        0.1617])
----------------------------------------
iter  0  stage  5  ep  0   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13,
13, 13, 13,
        13, 13, 13, 13, 13, 13,  0])
loss=  tensor(0.0154, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.9983,
6.9983, 6.9983, 6.9983, 6.9983, 6.9983, 6.4928, 6.0396, 5.6236,
        5.2347, 4.8653, 4.5103, 4.1659, 3.8293, 3.4985, 3.1721, 2.8489, 2.5281,
        2.2092, 1.8915, 1.5749, 1.2590, 0.9437, 0.6289, 0.3143]) return=
119073.90192566637
probs of actions:  tensor([0.9995, 0.9991, 0.9992, 0.9993, 0.9993, 0.9995,
0.9990, 0.9990, 0.9994,
        0.9998, 0.9996, 0.9996, 0.9998, 0.9999, 0.9994, 0.9993, 0.9998, 0.9994,
        0.9995, 0.9996, 0.9999, 1.0000, 0.9999, 0.9931, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4943, 0.5550, 0.5302, 0.5120, 0.4986, 0.4886, 0.4811,
0.4756, 0.4715,
        0.4684, 0.4661, 0.4644, 0.4631, 0.4621, 0.4614, 0.4608, 0.4604, 0.4601,
        0.4599, 0.4597, 0.4596, 0.4595, 0.4594, 0.4594, 0.4762])
finalReturns:  tensor([2.3190, 2.3359, 2.3080, 2.2483, 2.1658, 2.0668, 1.9557,
1.8357, 1.7092,
        1.5779, 1.4429, 1.3053, 1.1657, 1.0245, 0.8823, 0.7392, 0.5955, 0.4513,
        0.3067, 0.1619])
----------------------------------------
iter  0  stage  4  ep  0   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13,
13, 13, 13,
        13, 13, 13, 13, 13, 13,  0])
loss=  tensor(0.0175, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.3486,
7.3486, 7.3486, 7.3486, 7.3486, 6.8331, 6.3728, 5.9518, 5.5591,
        5.1870, 4.8301, 4.4842, 4.1465, 3.8150, 3.4879, 3.1643, 2.8432, 2.5240,
        2.2061, 1.8894, 1.5734, 1.2580, 0.9431, 0.6285, 0.3142]) return=
119073.90192566637
probs of actions:  tensor([0.9995, 0.9991, 0.9992, 0.9993, 0.9993, 0.9995,
0.9990, 0.9990, 0.9994,
        0.9998, 0.9996, 0.9996, 0.9998, 0.9999, 0.9994, 0.9993, 0.9998, 0.9994,
        0.9995, 0.9996, 0.9999, 1.0000, 0.9999, 0.9931, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4943, 0.5550, 0.5302, 0.5120, 0.4986, 0.4886, 0.4811,
0.4756, 0.4715,
        0.4684, 0.4661, 0.4644, 0.4631, 0.4621, 0.4614, 0.4608, 0.4604, 0.4601,
        0.4599, 0.4597, 0.4596, 0.4595, 0.4594, 0.4594, 0.4762])
finalReturns:  tensor([2.4672, 2.4841, 2.4559, 2.3957, 2.3128, 2.2134, 2.1020,
1.9818, 1.8551,
        1.7236, 1.5885, 1.4507, 1.3110, 1.1698, 1.0275, 0.8844, 0.7407, 0.5965,
        0.4519, 0.3071, 0.1621])
```

```
---------------------------------------
iter  0  stage  3  ep  0   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13,
13, 13, 13,
        13, 13, 13, 13, 13, 13,  0])
loss=  tensor(0.0196, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.7112,
7.7112, 7.7112, 7.7112, 7.1822, 6.7124, 6.2846, 5.8870, 5.5113,
        5.1517, 4.8038, 4.4647, 4.1321, 3.8043, 3.4800, 3.1585, 2.8389, 2.5208,
        2.2039, 1.8878, 1.5723, 1.2573, 0.9426, 0.6282, 0.3140]) return=
119073.90192566637
probs of actions:  tensor([0.9995, 0.9991, 0.9992, 0.9993, 0.9993, 0.9995,
0.9990, 0.9990, 0.9994,
        0.9998, 0.9996, 0.9996, 0.9998, 0.9999, 0.9994, 0.9993, 0.9998, 0.9994,
        0.9995, 0.9996, 0.9999, 1.0000, 0.9999, 0.9931, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4943, 0.5550, 0.5302, 0.5120, 0.4986, 0.4886, 0.4811,
0.4756, 0.4715,
        0.4684, 0.4661, 0.4644, 0.4631, 0.4621, 0.4614, 0.4608, 0.4604, 0.4601,
        0.4599, 0.4597, 0.4596, 0.4595, 0.4594, 0.4594, 0.4762])
finalReturns:  tensor([2.6167, 2.6336, 2.6048, 2.5441, 2.4606, 2.3606, 2.2488,
2.1282, 2.0012,
        1.8695, 1.7343, 1.5964, 1.4566, 1.3153, 1.1730, 1.0298, 0.8860, 0.7418,
        0.5972, 0.4524, 0.3073, 0.1622])
---------------------------------------
iter  0  stage  2  ep  0   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13,
13, 13, 13,
        13, 13, 13, 13, 13, 13,  0])
loss=  tensor(0.0223, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.0905,
8.0905, 8.0905, 7.5434, 7.0607, 6.6237, 6.2194, 5.8389, 5.4757,
        5.1253, 4.7843, 4.4502, 4.1213, 3.7963, 3.4741, 3.1541, 2.8357, 2.5185,
        2.2022, 1.8866, 1.5715, 1.2567, 0.9423, 0.6280, 0.3140]) return=
119073.90192566637
probs of actions:  tensor([0.9995, 0.9991, 0.9992, 0.9993, 0.9993, 0.9995,
0.9990, 0.9990, 0.9994,
        0.9998, 0.9996, 0.9996, 0.9998, 0.9999, 0.9994, 0.9993, 0.9998, 0.9994,
        0.9995, 0.9996, 0.9999, 1.0000, 0.9999, 0.9931, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4943, 0.5550, 0.5302, 0.5120, 0.4986, 0.4886, 0.4811,
0.4756, 0.4715,
        0.4684, 0.4661, 0.4644, 0.4631, 0.4621, 0.4614, 0.4608, 0.4604, 0.4601,
        0.4599, 0.4597, 0.4596, 0.4595, 0.4594, 0.4594, 0.4762])
finalReturns:  tensor([2.7675, 2.7844, 2.7551, 2.6936, 2.6093, 2.5086, 2.3962,
2.2752, 2.1478,
        2.0158, 1.8803, 1.7423, 1.6023, 1.4610, 1.3185, 1.1753, 1.0315, 0.8872,
        0.7426, 0.5978, 0.4527, 0.3075, 0.1623])
---------------------------------------
iter  0  stage  1  ep  0   adversary:  AdversaryModes.fight_lb_132
```

```
  actions:  tensor([13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13,
13, 13, 13,
        13, 13, 13, 13, 13, 13,  0])
loss=  tensor(0.0253, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.4928,
8.4928, 7.9208, 7.4207, 6.9713, 6.5581, 6.1711, 5.8032, 5.4492,
        5.1056, 4.7696, 4.4393, 4.1132, 3.7903, 3.4697, 3.1508, 2.8333, 2.5167,
        2.2009, 1.8857, 1.5708, 1.2563, 0.9420, 0.6279, 0.3139]) return=
119073.90192566637
probs of actions:  tensor([0.9995, 0.9991, 0.9992, 0.9993, 0.9993, 0.9995,
0.9990, 0.9990, 0.9995,
        0.9998, 0.9996, 0.9996, 0.9998, 0.9999, 0.9994, 0.9993, 0.9998, 0.9994,
        0.9995, 0.9996, 0.9999, 1.0000, 0.9999, 0.9931, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4943, 0.5550, 0.5302, 0.5120, 0.4986, 0.4886, 0.4811,
0.4756, 0.4715,
        0.4684, 0.4661, 0.4644, 0.4631, 0.4621, 0.4614, 0.4608, 0.4604, 0.4601,
        0.4599, 0.4597, 0.4596, 0.4595, 0.4594, 0.4594, 0.4762])
finalReturns:  tensor([2.9203, 2.9372, 2.9071, 2.8445, 2.7591, 2.6576, 2.5444,
2.4227, 2.2948,
        2.1624, 2.0267, 1.8884, 1.7483, 1.6068, 1.4642, 1.3209, 1.1771, 1.0328,
        0.8881, 0.7432, 0.5982, 0.4530, 0.3077, 0.1623])
----------------------------------------
iter  0  stage  0  ep  0   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13,
13, 13, 13,
        13, 13, 13, 13, 13, 13,  0])
loss=  tensor(0.0272, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.8318,
8.3206, 7.7967, 7.3305, 6.9053, 6.5095, 6.1352, 5.7766, 5.4295,
        5.0909, 4.7587, 4.4312, 4.1072, 3.7858, 3.4663, 3.1484, 2.8315, 2.5154,
        2.2000, 1.8850, 1.5704, 1.2560, 0.9418, 0.6278, 0.3138]) return=
119073.90192566637
probs of actions:  tensor([0.9995, 0.9991, 0.9992, 0.9993, 0.9993, 0.9995,
0.9990, 0.9990, 0.9995,
        0.9998, 0.9996, 0.9996, 0.9998, 0.9999, 0.9994, 0.9993, 0.9998, 0.9994,
        0.9995, 0.9996, 0.9999, 1.0000, 0.9999, 0.9931, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4943, 0.5550, 0.5302, 0.5120, 0.4986, 0.4886, 0.4811,
0.4756, 0.4715,
        0.4684, 0.4661, 0.4644, 0.4631, 0.4621, 0.4614, 0.4608, 0.4604, 0.4601,
        0.4599, 0.4597, 0.4596, 0.4595, 0.4594, 0.4594, 0.4762])
finalReturns:  tensor([3.0756, 3.0925, 3.0613, 2.9973, 2.9105, 2.8077, 2.6935,
2.5710, 2.4424,
        2.3095, 2.1734, 2.0348, 1.8944, 1.7528, 1.6101, 1.4667, 1.3227, 1.1784,
        1.0337, 0.8888, 0.7437, 0.5985, 0.4532, 0.3078, 0.1624])
0,[1e-05,1][1, 10000, 1, 1],1682999908 saved
[674563, 'tensor([0., 0., 0., 0., 0., 0., 1., 0., 0., 0., 0., 0.])',
119073.90192566637, 95284.79046306085, 0.027228858321905136, 1e-05, 1, 0,
'tensor([13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13,
```

```
13,\n        13, 13, 13, 13, 13, 13,  0])', '[1.   1.   1.   1.   1.   1.   1.
1.   1.   1.   1.   1.   1.   1.\n 1.   1.   1.   1.   1.   1.   1.   1.   1.
0.99 1.  ]', '0,[1e-05,1][1, 10000, 1, 1],1682999908', 25, 50,
158873.58343048888, 178559.0269066819, 70287.27661109495, 135436.67200000002,
132554.6666666667, 119073.90192566635, 119073.90192566635, 125827.28267117329,
125827.28267117329, 81831.91688437786, 119073.90192566635, 125827.28267117329]
policy reset
----------------------------------------
iter  1  stage  24  ep  99999   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
0, 2, 0, 0,
        0])
loss=  tensor(-0., grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.4718,
0.4718, 0.4718, 0.4718, 0.4718, 0.4718, 0.4718, 0.4718, 0.4718, 0.4718,
        0.4718, 0.4718, 0.4718, 0.4718, 0.4718, 0.4718, 0.4718, 0.4718, 0.4718,
        0.4718, 0.4718, 0.4718, 0.4718, 0.4718, 0.4718, 0.4718]) return=
131596.28685975395
probs of actions:  tensor([0.9367, 0.9428, 0.9187, 0.9468, 0.9307, 0.9152,
0.9421, 0.9478, 0.9220,
        0.9468, 0.9338, 0.9398, 0.9479, 0.0224, 0.9485, 0.9489, 0.9547, 0.9500,
        0.9286, 0.9317, 0.9214, 0.0105, 0.9444, 0.9322, 0.9943],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5112, 0.5238, 0.5334, 0.5406, 0.5460, 0.5501, 0.5532,
0.5555, 0.5573,
        0.5586, 0.5595, 0.5603, 0.5608, 0.5612, 0.5653, 0.5386, 0.5190, 0.5046,
        0.4938, 0.4859, 0.4800, 0.4751, 0.4791, 0.4749, 0.4718])
finalReturns:  tensor([0.])
----------------------------------------
iter  1  stage  23  ep  99999   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([ 9, 10,  4,  9,  9,  0,  0,  7, 10, 10,  9,  0, 11,  9,  7,
0,  9,  8,
         7,  8,  8,  9,  9,  9,  0])
loss=  tensor(0.0078, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.7901,
0.7901, 0.7901, 0.7901, 0.7901, 0.7901, 0.7901, 0.7901, 0.7901,
        0.7901, 0.7901, 0.7901, 0.7901, 0.7901, 0.7901, 0.7901, 0.7901, 0.7901,
        0.7901, 0.7901, 0.7901, 0.7901, 0.7901, 0.7901, 0.3823]) return=
106695.37744149746
probs of actions:  tensor([0.4556, 0.1081, 0.0112, 0.4412, 0.4102, 0.1664,
0.2821, 0.0526, 0.1414,
        0.1179, 0.4751, 0.2725, 0.0115, 0.4704, 0.0645, 0.2235, 0.4843, 0.0924,
        0.0722, 0.1001, 0.0992, 0.3777, 0.4121, 0.6827, 0.9964],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.5469, 0.5947, 0.5226, 0.4935, 0.4803, 0.4345,
0.3968, 0.3897,
        0.3977, 0.4057, 0.4152, 0.3756, 0.3937, 0.4012, 0.4030, 0.3707, 0.3823,
        0.3882, 0.3869, 0.3901, 0.3909, 0.3959, 0.3997, 0.4106])
finalReturns:  tensor([0.0202, 0.0283])
----------------------------------------
```

```
iter  1  stage  22  ep  99999   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([10, 10,  9, 10,  9, 10, 10,  9, 10, 10, 16, 10,  9, 10,  9,
10, 10, 10,
        10, 11, 16,  9, 11, 10,  0])
loss=  tensor(0.2371, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.2213,
1.2213, 1.2213, 1.2213, 1.2213, 1.2213, 1.2213, 1.2213, 1.2213,
        1.2213, 1.2213, 1.2213, 1.2213, 1.2213, 1.2213, 1.2213, 1.2213, 1.2213,
        1.2213, 1.2213, 1.2213, 1.2213, 1.2213, 0.7816, 0.3779]) return=
113419.76902528579
probs of actions:  tensor([0.5822, 0.5528, 0.2460, 0.5698, 0.2272, 0.5105,
0.5893, 0.2325, 0.5383,
        0.5680, 0.0183, 0.5190, 0.2591, 0.5497, 0.2466, 0.5711, 0.5934, 0.5512,
        0.5538, 0.0713, 0.0181, 0.2689, 0.0841, 0.4928, 0.9997],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5012, 0.5506, 0.5911, 0.5393, 0.5089, 0.4798, 0.4634,
0.4531, 0.4389,
        0.4330, 0.4131, 0.4455, 0.4398, 0.4290, 0.4276, 0.4199, 0.4189, 0.4181,
        0.4175, 0.4150, 0.4044, 0.4407, 0.4276, 0.4295, 0.4360])
finalReturns:  tensor([0.0718, 0.0839, 0.0582])
----------------------------------------
iter  1  stage  21  ep  99999   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([20, 16, 20, 20, 16, 16, 20, 20, 20, 20, 20, 10, 14, 20, 16,
16, 20, 16,
        16, 20, 16, 16, 16, 14,  0])
loss=  tensor(0.8845, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.1640,
1.1640, 1.1640, 1.1640, 1.1640, 1.1640, 1.1640, 1.1640, 1.1640,
        1.1640, 1.1640, 1.1640, 1.1640, 1.1640, 1.1640, 1.1640, 1.1640, 1.1640,
        1.1640, 1.1640, 1.1640, 1.1640, 0.8126, 0.5105, 0.2430]) return=
91897.17702893353
probs of actions:  tensor([0.5741, 0.2196, 0.5670, 0.5593, 0.1998, 0.2072,
0.5479, 0.5890, 0.5455,
        0.5652, 0.5592, 0.0451, 0.0846, 0.5561, 0.2024, 0.2001, 0.5917, 0.1945,
        0.2057, 0.5518, 0.2183, 0.1572, 0.1927, 0.1121, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4712, 0.5731, 0.4861, 0.4486, 0.4358, 0.4027, 0.3643,
0.3593, 0.3555,
        0.3527, 0.3506, 0.3791, 0.3378, 0.3062, 0.3299, 0.3251, 0.3071, 0.3306,
        0.3256, 0.3075, 0.3309, 0.3259, 0.3221, 0.3252, 0.3369])
finalReturns:  tensor([0.1460, 0.1716, 0.1516, 0.0939])
----------------------------------------
iter  1  stage  20  ep  99999   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,
20, 20, 20,
        20, 20, 20, 20, 20, 20,  0])
loss=  tensor(0.0227, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.4830,
1.4830, 1.4830, 1.4830, 1.4830, 1.4830, 1.4830, 1.4830, 1.4830,
        1.4830, 1.4830, 1.4830, 1.4830, 1.4830, 1.4830, 1.4830, 1.4830, 1.4830,
        1.4830, 1.4830, 1.4830, 1.0978, 0.7724, 0.4884, 0.2338]) return=
```

```
95926.91902269641
probs of actions:  tensor([0.9743, 0.9730, 0.9708, 0.9723, 0.9681, 0.9578,
0.9686, 0.9745, 0.9643,
        0.9707, 0.9698, 0.9638, 0.9648, 0.9697, 0.9709, 0.9736, 0.9771, 0.9706,
        0.9658, 0.9669, 0.9809, 0.9906, 0.9810, 0.9655, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4712, 0.5587, 0.5007, 0.4591, 0.4290, 0.4071, 0.3910,
0.3791, 0.3703,
        0.3637, 0.3588, 0.3552, 0.3525, 0.3505, 0.3489, 0.3478, 0.3470, 0.3463,
        0.3458, 0.3455, 0.3452, 0.3450, 0.3449, 0.3447, 0.3847])
finalReturns:  tensor([0.2814, 0.3214, 0.3019, 0.2410, 0.1509])
----------------------------------------
iter  1  stage  19  ep  75221   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([16, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,
20, 20, 20,
        20, 20, 20, 20, 20, 20,  0])
loss=  tensor(0.0016, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.2476,
2.2476, 2.2476, 2.2476, 2.2476, 2.2476, 2.2476, 2.2476, 2.2476, 2.2476,
        2.2476, 2.2476, 2.2476, 2.2476, 2.2476, 2.2476, 2.2476, 2.2476, 2.2476,
        2.2476, 2.2476, 1.7571, 1.3343, 0.9589, 0.6172, 0.2998]) return=
116915.56065675804
probs of actions:  tensor([0.0013, 0.9986, 0.9984, 0.9986, 0.9983, 0.9976,
0.9984, 0.9988, 0.9981,
        0.9985, 0.9985, 0.9982, 0.9982, 0.9985, 0.9986, 0.9988, 0.9990, 0.9986,
        0.9983, 0.9991, 0.9993, 0.9998, 0.9994, 0.9986, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4856, 0.5433, 0.5192, 0.5015, 0.4884, 0.4786, 0.4714,
0.4660, 0.4620,
        0.4590, 0.4567, 0.4550, 0.4538, 0.4528, 0.4521, 0.4516, 0.4512, 0.4509,
        0.4507, 0.4505, 0.4504, 0.4503, 0.4502, 0.4502, 0.4901])
finalReturns:  tensor([0.4940, 0.5340, 0.5065, 0.4316, 0.3231, 0.1903])
----------------------------------------
iter  1  stage  18  ep  10555   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([16, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,
20, 20, 20,
        20, 20, 20, 20, 20, 20,  0])
loss=  tensor(0.0016, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.5351,
2.5351, 2.5351, 2.5351, 2.5351, 2.5351, 2.5351, 2.5351, 2.5351, 2.5351,
        2.5351, 2.5351, 2.5351, 2.5351, 2.5351, 2.5351, 2.5351, 2.5351, 2.5351,
        2.5351, 2.0444, 1.6215, 1.2460, 0.9042, 0.5868, 0.2870]) return=
116915.56065675804
probs of actions:  tensor([8.3469e-04, 9.9912e-01, 9.9896e-01, 9.9913e-01,
9.9889e-01, 9.9846e-01,
        9.9898e-01, 9.9922e-01, 9.9876e-01, 9.9908e-01, 9.9905e-01, 9.9887e-01,
        9.9881e-01, 9.9905e-01, 9.9910e-01, 9.9924e-01, 9.9937e-01, 9.9911e-01,
        9.9914e-01, 9.9956e-01, 9.9962e-01, 9.9994e-01, 9.9966e-01, 9.9913e-01,
        1.0000e+00], grad_fn=<ExpBackward0>)
rewards:  tensor([0.4856, 0.5433, 0.5192, 0.5015, 0.4884, 0.4786, 0.4714,
```

```
0.4660, 0.4620,
        0.4590, 0.4567, 0.4550, 0.4538, 0.4528, 0.4521, 0.4516, 0.4512, 0.4509,
        0.4507, 0.4505, 0.4504, 0.4503, 0.4502, 0.4502, 0.4901])
finalReturns:  tensor([0.6572, 0.6972, 0.6697, 0.5948, 0.4862, 0.3535, 0.2032])
----------------------------------------
iter  1  stage  17  ep  51   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,
20, 20, 20,
        20, 20, 20, 20, 20, 20,  0])
loss=  tensor(0.0027, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.1125,
2.1125, 2.1125, 2.1125, 2.1125, 2.1125, 2.1125, 2.1125, 2.1125, 2.1125,
        2.1125, 2.1125, 2.1125, 2.1125, 2.1125, 2.1125, 2.1125, 2.1125, 2.1125,
        1.7262, 1.4000, 1.1155, 0.8604, 0.6264, 0.4075, 0.1997]) return=
95926.91902269641
probs of actions:  tensor([0.9991, 0.9991, 0.9989, 0.9991, 0.9988, 0.9983,
0.9988, 0.9991, 0.9986,
        0.9989, 0.9989, 0.9986, 0.9986, 0.9989, 0.9989, 0.9991, 0.9992, 0.9990,
        0.9990, 0.9995, 0.9995, 0.9999, 0.9996, 0.9989, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4712, 0.5587, 0.5007, 0.4591, 0.4290, 0.4071, 0.3910,
0.3791, 0.3703,
        0.3637, 0.3588, 0.3552, 0.3525, 0.3505, 0.3489, 0.3478, 0.3470, 0.3463,
        0.3458, 0.3455, 0.3452, 0.3450, 0.3449, 0.3447, 0.3847])
finalReturns:  tensor([0.6896, 0.7296, 0.7100, 0.6490, 0.5588, 0.4478, 0.3219,
0.1849])
----------------------------------------
iter  1  stage  16  ep  0   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,
20, 20, 20,
        20, 20, 20, 20, 20, 20,  0])
loss=  tensor(0.0038, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.3083,
2.3083, 2.3083, 2.3083, 2.3083, 2.3083, 2.3083, 2.3083, 2.3083, 2.3083,
        2.3083, 2.3083, 2.3083, 2.3083, 2.3083, 2.3083, 2.3083, 2.3083, 1.9214,
        1.5947, 1.3099, 1.0547, 0.8205, 0.6015, 0.3936, 0.1938]) return=
95926.91902269641
probs of actions:  tensor([0.9991, 0.9991, 0.9989, 0.9991, 0.9988, 0.9983,
0.9988, 0.9991, 0.9986,
        0.9989, 0.9989, 0.9986, 0.9986, 0.9989, 0.9989, 0.9991, 0.9992, 0.9990,
        0.9990, 0.9995, 0.9995, 0.9999, 0.9996, 0.9989, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4712, 0.5587, 0.5007, 0.4591, 0.4290, 0.4071, 0.3910,
0.3791, 0.3703,
        0.3637, 0.3588, 0.3552, 0.3525, 0.3505, 0.3489, 0.3478, 0.3470, 0.3463,
        0.3458, 0.3455, 0.3452, 0.3450, 0.3449, 0.3447, 0.3847])
finalReturns:  tensor([0.8407, 0.8807, 0.8611, 0.8000, 0.7098, 0.5988, 0.4728,
0.3358, 0.1909])
----------------------------------------
iter  1  stage  15  ep  0   adversary:  AdversaryModes.fight_lb_132
```

```
  actions:  tensor([20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,
20, 20, 20,
        20, 20, 20, 20, 20, 20,  0])
loss=  tensor(0.0052, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.5005,
2.5005, 2.5005, 2.5005, 2.5005, 2.5005, 2.5005, 2.5005, 2.5005,
        2.5005, 2.5005, 2.5005, 2.5005, 2.5005, 2.5005, 2.5005, 2.1127, 1.7854,
        1.5002, 1.2447, 1.0103, 0.7912, 0.5831, 0.3833, 0.1894]) return=
95926.91902269641
probs of actions:  tensor([0.9991, 0.9991, 0.9989, 0.9991, 0.9988, 0.9983,
0.9988, 0.9991, 0.9986,
        0.9989, 0.9989, 0.9986, 0.9986, 0.9989, 0.9989, 0.9991, 0.9992, 0.9990,
        0.9990, 0.9995, 0.9995, 0.9999, 0.9996, 0.9989, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4712, 0.5587, 0.5007, 0.4591, 0.4290, 0.4071, 0.3910,
0.3791, 0.3703,
        0.3637, 0.3588, 0.3552, 0.3525, 0.3505, 0.3489, 0.3478, 0.3470, 0.3463,
        0.3458, 0.3455, 0.3452, 0.3450, 0.3449, 0.3447, 0.3847])
finalReturns:  tensor([0.9963, 1.0363, 1.0166, 0.9555, 0.8653, 0.7542, 0.6281,
0.4911, 0.3461,
        0.1952])
----------------------------------------
iter  1  stage  14  ep  37   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,
20, 20, 20,
        20, 20, 20, 20, 20, 20,  0])
loss=  tensor(0.0065, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.6904,
2.6904, 2.6904, 2.6904, 2.6904, 2.6904, 2.6904, 2.6904, 2.6904,
        2.6904, 2.6904, 2.6904, 2.6904, 2.6904, 2.6904, 2.6904, 2.3015, 1.9734, 1.6877,
        1.4317, 1.1971, 0.9777, 0.7696, 0.5696, 0.3756, 0.1862]) return=
95926.91902269641
probs of actions:  tensor([0.9992, 0.9991, 0.9990, 0.9991, 0.9988, 0.9983,
0.9989, 0.9991, 0.9986,
        0.9989, 0.9989, 0.9987, 0.9986, 0.9989, 0.9990, 0.9991, 0.9993, 0.9991,
        0.9990, 0.9995, 0.9995, 0.9999, 0.9996, 0.9990, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4712, 0.5587, 0.5007, 0.4591, 0.4290, 0.4071, 0.3910,
0.3791, 0.3703,
        0.3637, 0.3588, 0.3552, 0.3525, 0.3505, 0.3489, 0.3478, 0.3470, 0.3463,
        0.3458, 0.3455, 0.3452, 0.3450, 0.3449, 0.3447, 0.3847])
finalReturns:  tensor([1.1554, 1.1954, 1.1756, 1.1144, 1.0240, 0.9129, 0.7867,
0.6497, 0.5047,
        0.3538, 0.1985])
----------------------------------------
iter  1  stage  13  ep  122   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,
20, 20, 20,
        20, 20, 20, 20, 20, 20,  0])
loss=  tensor(0.0081, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.8792,
```

```
2.8792, 2.8792, 2.8792, 2.8792, 2.8792, 2.8792, 2.8792, 2.8792,
        2.8792, 2.8792, 2.8792, 2.8792, 2.8792, 2.4887, 2.1597, 1.8732, 1.6167,
        1.3817, 1.1621, 0.9537, 0.7536, 0.5595, 0.3700, 0.1837]) return=
95926.91902269641
probs of actions:  tensor([0.9991, 0.9991, 0.9989, 0.9991, 0.9988, 0.9983,
0.9988, 0.9991, 0.9985,
        0.9989, 0.9989, 0.9986, 0.9986, 0.9990, 0.9991, 0.9992, 0.9994, 0.9991,
        0.9990, 0.9994, 0.9995, 0.9999, 0.9996, 0.9989, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4712, 0.5587, 0.5007, 0.4591, 0.4290, 0.4071, 0.3910,
0.3791, 0.3703,
        0.3637, 0.3588, 0.3552, 0.3525, 0.3505, 0.3489, 0.3478, 0.3470, 0.3463,
        0.3458, 0.3455, 0.3452, 0.3450, 0.3449, 0.3447, 0.3847])
finalReturns:  tensor([1.3170, 1.3570, 1.3372, 1.2759, 1.1854, 1.0741, 0.9479,
0.8108, 0.6657,
        0.5147, 0.3594, 0.2009])
----------------------------------------
iter  1  stage  12  ep  263   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,
20, 20, 20,
        20, 20, 20, 20, 20, 20,  0])
loss=  tensor(0.0084, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.0679,
3.0679, 3.0679, 3.0679, 3.0679, 3.0679, 3.0679, 3.0679, 3.0679,
        3.0679, 3.0679, 3.0679, 3.0679, 2.6754, 2.3450, 2.0575, 1.8003, 1.5648,
        1.3448, 1.1362, 0.9359, 0.7417, 0.5520, 0.3657, 0.1819]) return=
95926.91902269641
probs of actions:  tensor([0.9993, 0.9992, 0.9991, 0.9992, 0.9989, 0.9985,
0.9990, 0.9992, 0.9987,
        0.9990, 0.9990, 0.9988, 0.9990, 0.9992, 0.9993, 0.9994, 0.9995, 0.9993,
        0.9991, 0.9995, 0.9996, 0.9999, 0.9996, 0.9991, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4712, 0.5587, 0.5007, 0.4591, 0.4290, 0.4071, 0.3910,
0.3791, 0.3703,
        0.3637, 0.3588, 0.3552, 0.3525, 0.3505, 0.3489, 0.3478, 0.3470, 0.3463,
        0.3458, 0.3455, 0.3452, 0.3450, 0.3449, 0.3447, 0.3847])
finalReturns:  tensor([1.4808, 1.5208, 1.5008, 1.4393, 1.3487, 1.2373, 1.1109,
0.9737, 0.8286,
        0.6775, 0.5222, 0.3637, 0.2027])
----------------------------------------
iter  1  stage  11  ep  203   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,
20, 20, 20,
        20, 20, 20, 20, 20, 20,  0])
loss=  tensor(0.0104, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.2577,
3.2577, 3.2577, 3.2577, 3.2577, 3.2577, 3.2577, 3.2577, 3.2577,
        3.2577, 3.2577, 3.2577, 2.8625, 2.5301, 2.2414, 1.9833, 1.7471, 1.5266,
        1.3176, 1.1171, 0.9227, 0.7329, 0.5465, 0.3626, 0.1806]) return=
95926.91902269641
```

```
probs of actions:  tensor([0.9992, 0.9992, 0.9990, 0.9992, 0.9989, 0.9984,
0.9990, 0.9992, 0.9987,
        0.9990, 0.9990, 0.9990, 0.9990, 0.9993, 0.9993, 0.9995, 0.9995, 0.9993,
        0.9991, 0.9995, 0.9996, 0.9999, 0.9996, 0.9990, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4712, 0.5587, 0.5007, 0.4591, 0.4290, 0.4071, 0.3910,
0.3791, 0.3703,
        0.3637, 0.3588, 0.3552, 0.3525, 0.3505, 0.3489, 0.3478, 0.3470, 0.3463,
        0.3458, 0.3455, 0.3452, 0.3450, 0.3449, 0.3447, 0.3847])
finalReturns:  tensor([1.6463, 1.6863, 1.6661, 1.6044, 1.5136, 1.4020, 1.2755,
1.1381, 0.9929,
        0.8418, 0.6864, 0.5278, 0.3668, 0.2041])
----------------------------------------
iter  1  stage  10  ep  31   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,
20, 20, 20,
        20, 20, 20, 20, 20, 20,  0])
loss=  tensor(0.0128, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.4496,
3.4496, 3.4496, 3.4496, 3.4496, 3.4496, 3.4496, 3.4496, 3.4496, 3.4496,
        3.4496, 3.4496, 3.0507, 2.7159, 2.4254, 2.1660, 1.9289, 1.7078, 1.4984,
        1.2975, 1.1028, 0.9128, 0.7263, 0.5423, 0.3602, 0.1796]) return=
95926.91902269641
probs of actions:  tensor([0.9992, 0.9992, 0.9990, 0.9991, 0.9989, 0.9984,
0.9989, 0.9992, 0.9987,
        0.9990, 0.9990, 0.9990, 0.9990, 0.9993, 0.9993, 0.9995, 0.9995, 0.9993,
        0.9990, 0.9995, 0.9996, 0.9999, 0.9996, 0.9990, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4712, 0.5587, 0.5007, 0.4591, 0.4290, 0.4071, 0.3910,
0.3791, 0.3703,
        0.3637, 0.3588, 0.3552, 0.3525, 0.3505, 0.3489, 0.3478, 0.3470, 0.3463,
        0.3458, 0.3455, 0.3452, 0.3450, 0.3449, 0.3447, 0.3847])
finalReturns:  tensor([1.8132, 1.8532, 1.8329, 1.7709, 1.6798, 1.5679, 1.4412,
1.3037, 1.1583,
        1.0071, 0.8516, 0.6930, 0.5319, 0.3692, 0.2051])
----------------------------------------
iter  1  stage  9  ep  0   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,
20, 20, 20,
        20, 20, 20, 20, 20, 20,  0])
loss=  tensor(0.0154, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.6449,
3.6449, 3.6449, 3.6449, 3.6449, 3.6449, 3.6449, 3.6449, 3.6449, 3.6449,
        3.6449, 3.2412, 2.9030, 2.6101, 2.3491, 2.1108, 1.8889, 1.6788, 1.4775,
        1.2825, 1.0922, 0.9055, 0.7214, 0.5392, 0.3585, 0.1788]) return=
95926.91902269641
probs of actions:  tensor([0.9992, 0.9992, 0.9990, 0.9991, 0.9989, 0.9984,
0.9989, 0.9992, 0.9987,
        0.9990, 0.9990, 0.9990, 0.9990, 0.9993, 0.9993, 0.9995, 0.9995, 0.9993,
        0.9990, 0.9995, 0.9996, 0.9999, 0.9996, 0.9990, 1.0000],
```

```
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4712, 0.5587, 0.5007, 0.4591, 0.4290, 0.4071, 0.3910,
0.3791, 0.3703,
        0.3637, 0.3588, 0.3552, 0.3525, 0.3505, 0.3489, 0.3478, 0.3470, 0.3463,
        0.3458, 0.3455, 0.3452, 0.3450, 0.3449, 0.3447, 0.3847])
finalReturns:  tensor([1.9816, 2.0216, 2.0010, 1.9386, 1.8471, 1.7349, 1.6079,
1.4702, 1.3246,
        1.1733, 1.0177, 0.8590, 0.6979, 0.5351, 0.3709, 0.2058])
----------------------------------------
iter  1  stage  8  ep  4590   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,
20, 20, 20,
        20, 20, 20, 20, 20, 20,  0])
loss=  tensor(0.0115, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.8453,
3.8453, 3.8453, 3.8453, 3.8453, 3.8453, 3.8453, 3.8453, 3.8453, 3.8453,
        3.4351, 3.0924, 2.7964, 2.5332, 2.2933, 2.0702, 1.8593, 1.6573, 1.4619,
        1.2713, 1.0843, 0.9000, 0.7177, 0.5369, 0.3571, 0.1783]) return=
95926.91902269641
probs of actions:  tensor([0.9994, 0.9993, 0.9992, 0.9993, 0.9991, 0.9987,
0.9991, 0.9993, 0.9990,
        0.9998, 0.9993, 0.9994, 0.9992, 0.9998, 0.9996, 0.9998, 0.9997, 0.9994,
        0.9992, 0.9996, 0.9997, 1.0000, 0.9997, 0.9992, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4712, 0.5587, 0.5007, 0.4591, 0.4290, 0.4071, 0.3910,
0.3791, 0.3703,
        0.3637, 0.3588, 0.3552, 0.3525, 0.3505, 0.3489, 0.3478, 0.3470, 0.3463,
        0.3458, 0.3455, 0.3452, 0.3450, 0.3449, 0.3447, 0.3847])
finalReturns:  tensor([2.1515, 2.1915, 2.1704, 2.1076, 2.0155, 1.9029, 1.7756,
1.6375, 1.4917,
        1.3402, 1.1845, 1.0256, 0.8645, 0.7016, 0.5374, 0.3722, 0.2064])
----------------------------------------
iter  1  stage  7  ep  0   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,
20, 20, 20,
        20, 20, 20, 20, 20, 20,  0])
loss=  tensor(0.0136, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.0530,
4.0530, 4.0530, 4.0530, 4.0530, 4.0530, 4.0530, 4.0530, 3.6339,
        3.2852, 2.9850, 2.7188, 2.4768, 2.2522, 2.0402, 1.8374, 1.6413, 1.4503,
        1.2630, 1.0784, 0.8959, 0.7149, 0.5351, 0.3562, 0.1778]) return=
95926.91902269641
probs of actions:  tensor([0.9994, 0.9993, 0.9992, 0.9993, 0.9991, 0.9987,
0.9991, 0.9993, 0.9990,
        0.9998, 0.9993, 0.9994, 0.9992, 0.9998, 0.9996, 0.9998, 0.9997, 0.9994,
        0.9992, 0.9996, 0.9997, 1.0000, 0.9997, 0.9992, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4712, 0.5587, 0.5007, 0.4591, 0.4290, 0.4071, 0.3910,
0.3791, 0.3703,
        0.3637, 0.3588, 0.3552, 0.3525, 0.3505, 0.3489, 0.3478, 0.3470, 0.3463,
```

```
        0.3458, 0.3455, 0.3452, 0.3450, 0.3449, 0.3447, 0.3847])
finalReturns:  tensor([2.3229, 2.3629, 2.3413, 2.2778, 2.1851, 2.0719, 1.9440,
1.8056, 1.6595,
        1.5077, 1.3518, 1.1928, 1.0315, 0.8686, 0.7043, 0.5391, 0.3732, 0.2068])
----------------------------------------
iter  1  stage  6  ep  0   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,
20, 20, 20,
        20, 20, 20, 20, 20, 20,  0])
loss=  tensor(0.0162, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.2707,
4.2707, 4.2707, 4.2707, 4.2707, 4.2707, 4.2707, 3.8397, 3.4828,
        3.1770, 2.9068, 2.6620, 2.4353, 2.2218, 2.0179, 1.8210, 1.6294, 1.4416,
        1.2567, 1.0740, 0.8928, 0.7129, 0.5338, 0.3554, 0.1775]) return=
95926.91902269641
probs of actions:  tensor([0.9994, 0.9993, 0.9992, 0.9993, 0.9991, 0.9987,
0.9991, 0.9993, 0.9990,
        0.9998, 0.9993, 0.9994, 0.9992, 0.9998, 0.9996, 0.9998, 0.9997, 0.9994,
        0.9992, 0.9996, 0.9997, 1.0000, 0.9997, 0.9992, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4712, 0.5587, 0.5007, 0.4591, 0.4290, 0.4071, 0.3910,
0.3791, 0.3703,
        0.3637, 0.3588, 0.3552, 0.3525, 0.3505, 0.3489, 0.3478, 0.3470, 0.3463,
        0.3458, 0.3455, 0.3452, 0.3450, 0.3449, 0.3447, 0.3847])
finalReturns:  tensor([2.4962, 2.5362, 2.5140, 2.4495, 2.3560, 2.2420, 2.1134,
1.9745, 1.8279,
        1.6758, 1.5196, 1.3605, 1.1990, 1.0359, 0.8716, 0.7064, 0.5404, 0.3740,
        0.2071])
----------------------------------------
iter  1  stage  5  ep  9276   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([20, 20, 20, 20, 16, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,
20, 20, 20,
        20, 20, 20, 20, 20, 20,  0])
loss=  tensor(0.0145, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.4575,
4.4575, 4.4575, 4.4575, 4.4575, 4.4575, 4.0237, 3.6649, 3.3577,
        3.0866, 2.8411, 2.6139, 2.4000, 2.1958, 1.9988, 1.8070, 1.6192, 1.4342,
        1.2514, 1.0702, 0.8902, 0.7111, 0.5327, 0.3548, 0.1772]) return=
95556.53901947841
probs of actions:  tensor([9.9949e-01, 9.9946e-01, 9.9936e-01, 9.9944e-01,
7.3200e-04, 9.9900e-01,
        9.9951e-01, 9.9947e-01, 9.9918e-01, 9.9991e-01, 9.9956e-01, 9.9956e-01,
        9.9941e-01, 9.9991e-01, 9.9975e-01, 9.9983e-01, 9.9981e-01, 9.9959e-01,
        9.9938e-01, 9.9969e-01, 9.9974e-01, 1.0000e+00, 9.9982e-01, 9.9942e-01,
        1.0000e+00], grad_fn=<ExpBackward0>)
rewards:  tensor([0.4712, 0.5587, 0.5007, 0.4591, 0.4434, 0.3938, 0.3812,
0.3718, 0.3649,
        0.3597, 0.3559, 0.3530, 0.3508, 0.3492, 0.3480, 0.3471, 0.3464, 0.3459,
        0.3455, 0.3453, 0.3450, 0.3449, 0.3448, 0.3447, 0.3846])
finalReturns:  tensor([2.6650, 2.7050, 2.6826, 2.6180, 2.5242, 2.4100, 2.2813,
```

```
2.1422, 1.9956,
        1.8434, 1.6872, 1.5279, 1.3665, 1.2034, 1.0390, 0.8738, 0.7078, 0.5413,
        0.3745, 0.2074])
---------------------------------------
iter  1  stage  4  ep  0   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,
20, 20, 20,
        20, 20, 20, 20, 20, 20,  0])
loss=  tensor(0.0170, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.7526,
4.7526, 4.7526, 4.7526, 4.7526, 4.2835, 3.9008, 3.5771, 3.2943,
        3.0405, 2.8074, 2.5891, 2.3818, 2.1824, 1.9889, 1.7998, 1.6138, 1.4303,
        1.2486, 1.0682, 0.8888, 0.7102, 0.5321, 0.3544, 0.1771]) return=
95926.91902269641
probs of actions:  tensor([0.9995, 0.9995, 0.9994, 0.9994, 0.9993, 0.9990,
0.9995, 0.9995, 0.9992,
        0.9999, 0.9996, 0.9996, 0.9994, 0.9999, 0.9998, 0.9998, 0.9998, 0.9996,
        0.9994, 0.9997, 0.9997, 1.0000, 0.9998, 0.9994, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4712, 0.5587, 0.5007, 0.4591, 0.4290, 0.4071, 0.3910,
0.3791, 0.3703,
        0.3637, 0.3588, 0.3552, 0.3525, 0.3505, 0.3489, 0.3478, 0.3470, 0.3463,
        0.3458, 0.3455, 0.3452, 0.3450, 0.3449, 0.3447, 0.3847])
finalReturns:  tensor([2.8504, 2.8904, 2.8660, 2.7988, 2.7025, 2.5860, 2.4554,
2.3148, 2.1670,
        2.0138, 1.8569, 1.6971, 1.5352, 1.3718, 1.2072, 1.0417, 0.8756, 0.7091,
        0.5421, 0.3749, 0.2075])
---------------------------------------
iter  1  stage  3  ep  0   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,
20, 20, 20,
        20, 20, 20, 20, 20, 20,  0])
loss=  tensor(0.0192, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.0293,
5.0293, 5.0293, 5.0293, 4.5301, 4.1271, 3.7894, 3.4968, 3.2361,
        2.9979, 2.7761, 2.5661, 2.3648, 2.1699, 1.9797, 1.7930, 1.6089, 1.4267,
        1.2460, 1.0663, 0.8875, 0.7093, 0.5316, 0.3541, 0.1770]) return=
95926.91902269641
probs of actions:  tensor([0.9995, 0.9995, 0.9994, 0.9994, 0.9993, 0.9990,
0.9995, 0.9995, 0.9992,
        0.9999, 0.9996, 0.9996, 0.9994, 0.9999, 0.9998, 0.9998, 0.9998, 0.9996,
        0.9994, 0.9997, 0.9997, 1.0000, 0.9998, 0.9994, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4712, 0.5587, 0.5007, 0.4591, 0.4290, 0.4071, 0.3910,
0.3791, 0.3703,
        0.3637, 0.3588, 0.3552, 0.3525, 0.3505, 0.3489, 0.3478, 0.3470, 0.3463,
        0.3458, 0.3455, 0.3452, 0.3450, 0.3449, 0.3447, 0.3847])
finalReturns:  tensor([3.0328, 3.0728, 3.0468, 2.9775, 2.8790, 2.7607, 2.6286,
2.4867, 2.3378,
        2.1839, 2.0264, 1.8661, 1.7039, 1.5402, 1.3754, 1.2098, 1.0436, 0.8769,
```

```
        0.7099, 0.5427, 0.3753, 0.2077])
---------------------------------------
iter  1  stage  2  ep  0   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,
20, 20, 20,
        20, 20, 20, 20, 20, 20,  0])
loss=  tensor(0.0217, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.3425,
5.3425, 5.3425, 4.8018, 4.3708, 4.0140, 3.7081, 3.4380, 3.1931,
        2.9665, 2.7529, 2.5490, 2.3522, 2.1605, 1.9728, 1.7879, 1.6051, 1.4240,
        1.2440, 1.0650, 0.8866, 0.7087, 0.5311, 0.3539, 0.1769]) return=
95926.91902269641
probs of actions:  tensor([0.9995, 0.9995, 0.9994, 0.9994, 0.9993, 0.9990,
0.9995, 0.9995, 0.9992,
        0.9999, 0.9996, 0.9996, 0.9994, 0.9999, 0.9998, 0.9998, 0.9998, 0.9996,
        0.9994, 0.9997, 0.9997, 1.0000, 0.9998, 0.9994, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4712, 0.5587, 0.5007, 0.4591, 0.4290, 0.4071, 0.3910,
0.3791, 0.3703,
        0.3637, 0.3588, 0.3552, 0.3525, 0.3505, 0.3489, 0.3478, 0.3470, 0.3463,
        0.3458, 0.3455, 0.3452, 0.3450, 0.3449, 0.3447, 0.3847])
finalReturns:  tensor([3.2203, 3.2603, 3.2321, 3.1600, 3.0587, 2.9379, 2.8037,
2.6601, 2.5099,
        2.3549, 2.1966, 2.0357, 1.8730, 1.7090, 1.5439, 1.3781, 1.2117, 1.0450,
        0.8779, 0.7106, 0.5431, 0.3755, 0.2078])
---------------------------------------
iter  1  stage  1  ep  0   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,
20, 20, 20,
        20, 20, 20, 20, 20, 20,  0])
loss=  tensor(0.0241, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.7070,
5.7070, 5.1084, 4.6387, 4.2556, 3.9315, 3.6486, 3.3946, 3.1614,
        2.9431, 2.7357, 2.5362, 2.3427, 2.1535, 1.9676, 1.7841, 1.6023, 1.4219,
        1.2426, 1.0639, 0.8858, 0.7082, 0.5308, 0.3537, 0.1768]) return=
95926.91902269641
probs of actions:  tensor([0.9995, 0.9995, 0.9994, 0.9994, 0.9993, 0.9990,
0.9995, 0.9995, 0.9992,
        0.9999, 0.9996, 0.9996, 0.9994, 0.9999, 0.9998, 0.9998, 0.9998, 0.9996,
        0.9994, 0.9997, 0.9997, 1.0000, 0.9998, 0.9994, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4712, 0.5587, 0.5007, 0.4591, 0.4290, 0.4071, 0.3910,
0.3791, 0.3703,
        0.3637, 0.3588, 0.3552, 0.3525, 0.3505, 0.3489, 0.3478, 0.3470, 0.3463,
        0.3458, 0.3455, 0.3452, 0.3450, 0.3449, 0.3447, 0.3847])
finalReturns:  tensor([3.4144, 3.4544, 3.4234, 3.3474, 3.2424, 3.1183, 2.9813,
2.8354, 2.6835,
        2.5271, 2.3677, 2.2060, 2.0427, 1.8782, 1.7128, 1.5467, 1.3801, 1.2132,
        1.0460, 0.8786, 0.7111, 0.5434, 0.3757, 0.2079])
---------------------------------------
```

```
iter  1  stage  0  ep  0   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,
20, 20, 20,
        20, 20, 20, 20, 20, 20,  0])
loss=  tensor(0.0264, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.9753,
5.4641, 4.9403, 4.5206, 4.1715, 3.8710, 3.6046, 3.3625, 3.1378,
        2.9257, 2.7228, 2.5267, 2.3357, 2.1483, 1.9637, 1.7812, 1.6003, 1.4204,
        1.2415, 1.0631, 0.8853, 0.7078, 0.5306, 0.3536, 0.1767]) return=
95926.91902269641
probs of actions:  tensor([0.9995, 0.9995, 0.9994, 0.9994, 0.9993, 0.9990,
0.9995, 0.9995, 0.9992,
        0.9999, 0.9996, 0.9996, 0.9994, 0.9999, 0.9998, 0.9998, 0.9998, 0.9996,
        0.9994, 0.9997, 0.9997, 1.0000, 0.9998, 0.9994, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4712, 0.5587, 0.5007, 0.4591, 0.4290, 0.4071, 0.3910,
0.3791, 0.3703,
        0.3637, 0.3588, 0.3552, 0.3525, 0.3505, 0.3489, 0.3478, 0.3470, 0.3463,
        0.3458, 0.3455, 0.3452, 0.3450, 0.3449, 0.3447, 0.3847])
finalReturns:  tensor([3.6174, 3.6574, 3.6225, 3.5415, 3.4315, 3.3029, 3.1622,
3.0134, 2.8590,
        2.7009, 2.5400, 2.3772, 2.2131, 2.0479, 1.8820, 1.7156, 1.5488, 1.3817,
        1.2143, 1.0468, 0.8792, 0.7114, 0.5436, 0.3758, 0.2079])
0,[1e-05,1][1, 10000, 1, 1],1683019988 saved
[700369, 'tensor([0., 0., 0., 0., 0., 0., 1., 0., 0., 0., 0., 0.])',
95926.91902269641, 83461.56370671562, 0.026398321613669395, 1e-05, 1, 0,
'tensor([20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,
20,\n        20, 20, 20, 20, 20, 20,  0])', '[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.\n 1.]', '0,[1e-05,1][1, 10000, 1,
1],1683019988', 25, 50, 166678.00013679263, 199440.72542558872,
81175.85726043607, 135011.53866666666, 132074.72800000003, 95926.91902269641,
95926.91902269641, 113529.90192566635, 113488.54729640356, 94070.91886744411,
95926.91902269641, 113529.90192566635]
policy reset
----------------------------------------
iter  2  stage  24  ep  99999   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([0, 0, 0, 0, 0, 0, 0, 0, 5, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0,
        0])
loss=  tensor(-0., grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.4335,
0.4335, 0.4335, 0.4335, 0.4335, 0.4335, 0.4335, 0.4335, 0.4335,
        0.4335, 0.4335, 0.4335, 0.4335, 0.4335, 0.4335, 0.4335, 0.4335, 0.4335,
        0.4335, 0.4335, 0.4335, 0.4335, 0.4335, 0.4335, 0.4335]) return=
123480.07884867929
probs of actions:  tensor([0.8697, 0.8787, 0.8837, 0.8861, 0.9065, 0.8957,
0.9092, 0.9046, 0.0033,
        0.8968, 0.0717, 0.8905, 0.8930, 0.8939, 0.8652, 0.9085, 0.8924, 0.8947,
        0.8934, 0.8794, 0.8891, 0.8795, 0.8696, 0.8838, 0.9802],
       grad_fn=<ExpBackward0>)
```

```
rewards:  tensor([0.5112, 0.5238, 0.5334, 0.5406, 0.5460, 0.5501, 0.5532,
0.5555, 0.5548,
        0.5774, 0.5389, 0.5146, 0.4932, 0.4774, 0.4658, 0.4571, 0.4507, 0.4459,
        0.4423, 0.4396, 0.4376, 0.4361, 0.4350, 0.4342, 0.4335])
finalReturns:  tensor([0.])
----------------------------------------
iter  2  stage  23  ep  99999   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([11, 10, 11,  8,  9,  0,  0, 11, 10,  9,  0,  6,  0,  8,  6,
11,  9,  6,
         0,  4,  1,  6,  9, 12,  0])
loss=  tensor(0.0690, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.8430,
0.8430, 0.8430, 0.8430, 0.8430, 0.8430, 0.8430, 0.8430, 0.8430, 0.8430,
        0.8430, 0.8430, 0.8430, 0.8430, 0.8430, 0.8430, 0.8430, 0.8430, 0.8430,
        0.8430, 0.8430, 0.8430, 0.8430, 0.8430, 0.8430, 0.4124]) return=
113897.17236003252
probs of actions:  tensor([0.3751, 0.0684, 0.3539, 0.0686, 0.1139, 0.2947,
0.1850, 0.3508, 0.0738,
        0.1131, 0.1590, 0.0437, 0.2453, 0.0760, 0.0570, 0.3085, 0.1086, 0.0428,
        0.2804, 0.0067, 0.0272, 0.0505, 0.1165, 0.0645, 0.9907],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5544, 0.5332, 0.5284, 0.5081, 0.5061, 0.4672,
0.4270, 0.4449,
        0.4555, 0.4667, 0.4351, 0.4379, 0.4113, 0.4251, 0.4183, 0.4402, 0.4516,
        0.4502, 0.4251, 0.4223, 0.4059, 0.4079, 0.4162, 0.4518])
finalReturns:  tensor([0.0250, 0.0394])
----------------------------------------
iter  2  stage  22  ep  99999   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
0, 11, 11,
        11, 11, 11, 11, 11, 11,  0])
loss=  tensor(0.0053, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.3927,
1.3927, 1.3927, 1.3927, 1.3927, 1.3927, 1.3927, 1.3927, 1.3927, 1.3927,
        1.3927, 1.3927, 1.3927, 1.3927, 1.3927, 1.3927, 1.3927, 1.3927, 1.3927,
        1.3927, 1.3927, 1.3927, 1.3927, 1.3927, 1.3927, 0.8954, 0.4345]) return=
124264.57333060534
probs of actions:  tensor([0.9427, 0.9337, 0.9446, 0.9470, 0.9577, 0.9128,
0.9357, 0.9532, 0.9385,
        0.9313, 0.9501, 0.9403, 0.9314, 0.9393, 0.9421, 0.0346, 0.9357, 0.9234,
        0.9208, 0.9168, 0.9496, 0.9306, 0.9677, 0.9730, 0.9974],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.5051, 0.4545, 0.4637,
        0.4707, 0.4760, 0.4800, 0.4830, 0.4852, 0.4869, 0.5003])
finalReturns:  tensor([0.0797, 0.0918, 0.0658])
----------------------------------------
iter  2  stage  21  ep  99999   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
```

```
11, 11, 11,
        11, 11, 12, 11, 11, 11,  0])
loss=  tensor(0.0080, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.8357,
1.8357, 1.8357, 1.8357, 1.8357, 1.8357, 1.8357, 1.8357, 1.8357,
        1.8357, 1.8357, 1.8357, 1.8357, 1.8357, 1.8357, 1.8357, 1.8357, 1.8357,
        1.8357, 1.8357, 1.8357, 1.8357, 1.3278, 0.8593, 0.4193]) return=
125646.00419333146
probs of actions:  tensor([0.9796, 0.9785, 0.9838, 0.9819, 0.9860, 0.9745,
0.9773, 0.9863, 0.9773,
        0.9804, 0.9844, 0.9810, 0.9799, 0.9795, 0.9794, 0.9820, 0.9782, 0.9743,
        0.9818, 0.9728, 0.0102, 0.9734, 0.9912, 0.9842, 0.9991],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4923, 0.4899, 0.4957, 0.4948, 0.4941, 0.5057])
finalReturns:  tensor([0.1547, 0.1668, 0.1405, 0.0864])
----------------------------------------
iter  2  stage  20  ep  99999   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11, 11, 11,
        11, 11, 11, 11, 11, 11,  0])
loss=  tensor(0.0292, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.2295,
2.2295, 2.2295, 2.2295, 2.2295, 2.2295, 2.2295, 2.2295, 2.2295,
        2.2295, 2.2295, 2.2295, 2.2295, 2.2295, 2.2295, 2.2295, 2.2295, 2.2295,
        2.2295, 2.2295, 2.2295, 1.7252, 1.2592, 0.8210, 0.4030]) return=
125571.79187222246
probs of actions:  tensor([0.9680, 0.9685, 0.9766, 0.9720, 0.9780, 0.9648,
0.9629, 0.9802, 0.9639,
        0.9731, 0.9777, 0.9711, 0.9713, 0.9694, 0.9686, 0.9748, 0.9664, 0.9592,
        0.9771, 0.9606, 0.9766, 0.9474, 0.9886, 0.9610, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4923, 0.4922, 0.4922, 0.4921, 0.4921, 0.5042])
finalReturns:  tensor([0.2433, 0.2554, 0.2292, 0.1753, 0.1011])
----------------------------------------
iter  2  stage  19  ep  99999   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
11, 11, 11,
        11, 12, 12, 11, 11, 11,  0])
loss=  tensor(1.7093, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.6218,
2.6218, 2.6218, 2.6218, 2.6218, 2.6218, 2.6218, 2.6218, 2.6218,
        2.6218, 2.6218, 2.6218, 2.6218, 2.6218, 2.6218, 2.6218, 2.6218, 2.6218,
        2.6218, 2.6218, 2.6218, 2.1174, 1.6514, 1.2131, 0.7951, 0.3921]) return=
125731.65338382448
probs of actions:  tensor([0.8931, 0.8975, 0.9188, 0.9044, 0.9218, 0.8893,
```

```
0.8750, 0.9309, 0.8813,
        0.9124, 0.9239, 0.9041, 0.9058, 0.8980, 0.8968, 0.9130, 0.8879, 0.8648,
        0.9240, 0.1488, 0.0878, 0.8260, 0.9609, 0.8325, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5523, 0.5369, 0.5255, 0.5170, 0.5107, 0.5060,
0.5025, 0.4998,
        0.4979, 0.4964, 0.4953, 0.4945, 0.4939, 0.4934, 0.4930, 0.4928, 0.4926,
        0.4924, 0.4900, 0.4935, 0.4984, 0.4968, 0.4956, 0.5068])
finalReturns:  tensor([0.3593, 0.3737, 0.3463, 0.2861, 0.2073, 0.1147])
----------------------------------------
iter  2  stage  18  ep  99999   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([12, 12, 12, 12, 12, 11, 12, 11, 12, 12, 12, 12, 12, 12, 11,
11, 12, 12,
        12, 11, 12, 12, 12, 12,  0])
loss=  tensor(1.3323, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.8592,
2.8592, 2.8592, 2.8592, 2.8592, 2.8592, 2.8592, 2.8592, 2.8592,
        2.8592, 2.8592, 2.8592, 2.8592, 2.8592, 2.8592, 2.8592, 2.8592, 2.8592,
        2.8592, 2.3724, 1.9258, 1.5082, 1.1117, 0.7307, 0.3611]) return=
121796.3805407076
probs of actions:  tensor([0.7982, 0.7813, 0.7673, 0.7875, 0.7589, 0.2079,
0.8283, 0.2679, 0.8133,
        0.7479, 0.7430, 0.7761, 0.7679, 0.7932, 0.2071, 0.2261, 0.8062, 0.8390,
        0.7591, 0.1343, 0.8476, 0.8865, 0.7336, 0.9199, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.5537, 0.5337, 0.5188, 0.5079, 0.5020, 0.4901,
0.4887, 0.4802,
        0.4790, 0.4782, 0.4775, 0.4770, 0.4767, 0.4787, 0.4750, 0.4699, 0.4714,
        0.4724, 0.4755, 0.4703, 0.4716, 0.4726, 0.4734, 0.4883])
finalReturns:  tensor([0.4650, 0.4794, 0.4505, 0.3978, 0.3226, 0.2310, 0.1272])
----------------------------------------
iter  2  stage  17  ep  99999   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 11, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0642, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.2258,
3.2258, 3.2258, 3.2258, 3.2258, 3.2258, 3.2258, 3.2258, 3.2258,
        3.2258, 3.2258, 3.2258, 3.2258, 3.2258, 3.2258, 3.2258, 3.2258, 3.2258,
        2.7355, 2.2864, 1.8670, 1.4692, 1.0872, 0.7169, 0.3553]) return=
122213.81800409526
probs of actions:  tensor([0.9715, 0.9674, 0.9688, 0.9709, 0.9668, 0.9693,
0.9768, 0.9617, 0.9737,
        0.0387, 0.9627, 0.9676, 0.9658, 0.9710, 0.9709, 0.9687, 0.9726, 0.9801,
        0.9701, 0.9851, 0.9858, 0.9873, 0.9717, 0.9934, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.5537, 0.5337, 0.5188, 0.5079, 0.4997, 0.4936,
0.4891, 0.4857,
        0.4855, 0.4778, 0.4772, 0.4768, 0.4765, 0.4763, 0.4761, 0.4760, 0.4759,
        0.4758, 0.4758, 0.4757, 0.4757, 0.4757, 0.4757, 0.4900])
```

```
finalReturns:  tensor([0.5944, 0.6088, 0.5821, 0.5258, 0.4478, 0.3541, 0.2487,
0.1348])
----------------------------------------
iter  2  stage  16  ep  99999   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0178, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.5788,
3.5788, 3.5788, 3.5788, 3.5788, 3.5788, 3.5788, 3.5788, 3.5788, 3.5788,
        3.5788, 3.5788, 3.5788, 3.5788, 3.5788, 3.5788, 3.5788, 3.5788, 3.0878,
        2.6382, 2.2184, 1.8203, 1.4382, 1.0677, 0.7059, 0.3506]) return=
122329.26544005591
probs of actions:  tensor([0.9930, 0.9917, 0.9927, 0.9930, 0.9921, 0.9921,
0.9943, 0.9909, 0.9935,
        0.9904, 0.9909, 0.9920, 0.9915, 0.9928, 0.9930, 0.9926, 0.9944, 0.9954,
        0.9950, 0.9971, 0.9974, 0.9972, 0.9951, 0.9989, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.5537, 0.5337, 0.5188, 0.5079, 0.4997, 0.4936,
0.4891, 0.4857,
        0.4832, 0.4813, 0.4798, 0.4788, 0.4780, 0.4774, 0.4769, 0.4766, 0.4764,
        0.4762, 0.4760, 0.4759, 0.4758, 0.4758, 0.4757, 0.4901])
finalReturns:  tensor([0.7197, 0.7341, 0.7074, 0.6510, 0.5731, 0.4793, 0.3739,
0.2599, 0.1395])
----------------------------------------
iter  2  stage  15  ep  99999   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0061, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.9270,
3.9270, 3.9270, 3.9270, 3.9270, 3.9270, 3.9270, 3.9270, 3.9270, 3.9270,
        3.9270, 3.9270, 3.9270, 3.9270, 3.9270, 3.9270, 3.9270, 3.4356, 2.9858,
        2.5658, 2.1676, 1.7854, 1.4148, 1.0530, 0.6976, 0.3470]) return=
122329.26544005591
probs of actions:  tensor([0.9975, 0.9970, 0.9975, 0.9976, 0.9973, 0.9971,
0.9980, 0.9969, 0.9977,
        0.9966, 0.9968, 0.9972, 0.9970, 0.9974, 0.9976, 0.9985, 0.9987, 0.9989,
        0.9986, 0.9990, 0.9995, 0.9995, 0.9989, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.5537, 0.5337, 0.5188, 0.5079, 0.4997, 0.4936,
0.4891, 0.4857,
        0.4832, 0.4813, 0.4798, 0.4788, 0.4780, 0.4774, 0.4769, 0.4766, 0.4764,
        0.4762, 0.4760, 0.4759, 0.4758, 0.4758, 0.4757, 0.4901])
finalReturns:  tensor([0.8485, 0.8629, 0.8361, 0.7797, 0.7018, 0.6080, 0.5026,
0.3886, 0.2682,
        0.1431])
----------------------------------------
iter  2  stage  15  ep  117999   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
```

```
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0040, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.9270,
3.9270, 3.9270, 3.9270, 3.9270, 3.9270, 3.9270, 3.9270, 3.9270,
        3.9270, 3.9270, 3.9270, 3.9270, 3.9270, 3.9270, 3.9270, 3.4356, 2.9858,
        2.5658, 2.1676, 1.7854, 1.4148, 1.0530, 0.6976, 0.3470]) return=
122329.26544005591
probs of actions:  tensor([0.9982, 0.9978, 0.9982, 0.9983, 0.9981, 0.9979,
0.9985, 0.9978, 0.9983,
        0.9975, 0.9977, 0.9980, 0.9978, 0.9981, 0.9983, 0.9990, 0.9990, 0.9992,
        0.9994, 0.9993, 0.9997, 0.9997, 0.9993, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.5537, 0.5337, 0.5188, 0.5079, 0.4997, 0.4936,
0.4891, 0.4857,
        0.4832, 0.4813, 0.4798, 0.4788, 0.4780, 0.4774, 0.4769, 0.4766, 0.4764,
        0.4762, 0.4760, 0.4759, 0.4758, 0.4758, 0.4757, 0.4901])
finalReturns:  tensor([0.8485, 0.8629, 0.8361, 0.7797, 0.7018, 0.6080, 0.5026,
0.3886, 0.2682,
        0.1431])
----------------------------------------
iter  2  stage  14  ep  908   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([11, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0052, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.5483,
4.5483, 4.5483, 4.5483, 4.5483, 4.5483, 4.5483, 4.5483, 4.5483,
        4.5483, 4.5483, 4.5483, 4.5483, 4.5483, 4.5483, 4.0289, 3.5520, 3.1059,
        2.6822, 2.2749, 1.8797, 1.4934, 1.1138, 0.7392, 0.3682]) return=
127915.76530179875
probs of actions:  tensor([0.0017, 0.9979, 0.9983, 0.9983, 0.9982, 0.9980,
0.9986, 0.9979, 0.9985,
        0.9977, 0.9979, 0.9982, 0.9979, 0.9983, 0.9990, 0.9991, 0.9991, 0.9993,
        0.9994, 0.9993, 0.9997, 0.9997, 0.9994, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5500, 0.5383, 0.5296, 0.5232, 0.5183, 0.5147,
0.5120, 0.5100,
        0.5085, 0.5074, 0.5065, 0.5059, 0.5054, 0.5051, 0.5048, 0.5046, 0.5045,
        0.5043, 0.5043, 0.5042, 0.5041, 0.5041, 0.5041, 0.5185])
finalReturns:  tensor([1.0141, 1.0285, 1.0006, 0.9421, 0.8614, 0.7644, 0.6553,
0.5374, 0.4128,
        0.2834, 0.1503])
----------------------------------------
iter  2  stage  13  ep  3619   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([11, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0067, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.9158,
4.9158, 4.9158, 4.9158, 4.9158, 4.9158, 4.9158, 4.9158, 4.9158,
```

```
        4.9158, 4.9158, 4.9158, 4.9158, 4.9158, 4.3960, 3.9188, 3.4725, 3.0487,
        2.6413, 2.2460, 1.8597, 1.4800, 1.1054, 0.7344, 0.3662]) return=
127915.76530179875
probs of actions:  tensor([0.0018, 0.9979, 0.9983, 0.9983, 0.9982, 0.9980,
0.9986, 0.9979, 0.9984,
        0.9977, 0.9979, 0.9982, 0.9979, 0.9990, 0.9990, 0.9991, 0.9991, 0.9993,
        0.9994, 0.9993, 0.9997, 0.9997, 0.9994, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4991, 0.5500, 0.5383, 0.5296, 0.5232, 0.5183, 0.5147,
0.5120, 0.5100,
        0.5085, 0.5074, 0.5065, 0.5059, 0.5054, 0.5051, 0.5048, 0.5046, 0.5045,
        0.5043, 0.5043, 0.5042, 0.5041, 0.5041, 0.5041, 0.5185])
finalReturns:  tensor([1.1521, 1.1665, 1.1386, 1.0801, 0.9994, 0.9023, 0.7932,
0.6753, 0.5507,
        0.4213, 0.2882, 0.1523])
----------------------------------------
iter  2  stage  12  ep  18243   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0071, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.9612,
4.9612, 4.9612, 4.9612, 4.9612, 4.9612, 4.9612, 4.9612, 4.9612, 4.9612,
        4.9612, 4.9612, 4.9612, 4.9612, 4.4680, 4.0168, 3.5959, 3.1970, 2.8143,
        2.4434, 2.0812, 1.7256, 1.3749, 1.0278, 0.6833, 0.3409]) return=
122329.26544005591
probs of actions:  tensor([0.9984, 0.9981, 0.9984, 0.9985, 0.9983, 0.9981,
0.9987, 0.9981, 0.9985,
        0.9978, 0.9980, 0.9983, 0.9990, 0.9992, 0.9991, 0.9992, 0.9991, 0.9993,
        0.9999, 0.9993, 0.9997, 0.9999, 0.9995, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.5537, 0.5337, 0.5188, 0.5079, 0.4997, 0.4936,
0.4891, 0.4857,
        0.4832, 0.4813, 0.4798, 0.4788, 0.4780, 0.4774, 0.4769, 0.4766, 0.4764,
        0.4762, 0.4760, 0.4759, 0.4758, 0.4758, 0.4757, 0.4901])
finalReturns:  tensor([1.2484, 1.2628, 1.2360, 1.1795, 1.1015, 1.0077, 0.9022,
0.7882, 0.6677,
        0.5426, 0.4139, 0.2825, 0.1492])
----------------------------------------
iter  2  stage  11  ep  5010   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0066, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.3049,
5.3049, 5.3049, 5.3049, 5.3049, 5.3049, 5.3049, 5.3049, 5.3049, 5.3049,
        5.3049, 5.3049, 5.3049, 4.8106, 4.3587, 3.9372, 3.5379, 3.1548, 2.7837,
        2.4214, 2.0657, 1.7149, 1.3677, 1.0232, 0.6807, 0.3398]) return=
122329.26544005591
probs of actions:  tensor([0.9988, 0.9985, 0.9988, 0.9988, 0.9987, 0.9986,
```

```
0.9990, 0.9986, 0.9989,
        0.9984, 0.9985, 0.9990, 0.9993, 0.9995, 0.9994, 0.9994, 0.9994, 0.9995,
        1.0000, 0.9995, 0.9998, 0.9999, 0.9997, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.5537, 0.5337, 0.5188, 0.5079, 0.4997, 0.4936,
0.4891, 0.4857,
        0.4832, 0.4813, 0.4798, 0.4788, 0.4780, 0.4774, 0.4769, 0.4766, 0.4764,
        0.4762, 0.4760, 0.4759, 0.4758, 0.4758, 0.4757, 0.4901])
finalReturns:  tensor([1.3846, 1.3990, 1.3722, 1.3157, 1.2376, 1.1437, 1.0382,
0.9241, 0.8037,
        0.6785, 0.5498, 0.4184, 0.2851, 0.1503])
---------------------------------------
iter  2  stage  10  ep  12616   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0071, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.6490,
5.6490, 5.6490, 5.6490, 5.6490, 5.6490, 5.6490, 5.6490, 5.6490,
        5.6490, 5.6490, 5.1533, 4.7004, 4.2782, 3.8783, 3.4949, 3.1234, 2.7610,
        2.4051, 2.0541, 1.7068, 1.3623, 1.0198, 0.6788, 0.3390]) return=
122329.26544005591
probs of actions:  tensor([0.9989, 0.9987, 0.9989, 0.9990, 0.9989, 0.9987,
0.9991, 0.9987, 0.9990,
        0.9985, 0.9990, 0.9991, 0.9994, 0.9997, 0.9995, 0.9995, 0.9996, 0.9995,
        1.0000, 0.9996, 0.9998, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.5537, 0.5337, 0.5188, 0.5079, 0.4997, 0.4936,
0.4891, 0.4857,
        0.4832, 0.4813, 0.4798, 0.4788, 0.4780, 0.4774, 0.4769, 0.4766, 0.4764,
        0.4762, 0.4760, 0.4759, 0.4758, 0.4758, 0.4757, 0.4901])
finalReturns:  tensor([1.5217, 1.5361, 1.5093, 1.4527, 1.3745, 1.2806, 1.1751,
1.0610, 0.9405,
        0.8153, 0.6865, 0.5552, 0.4218, 0.2870, 0.1511])
---------------------------------------
iter  2  stage  9  ep  510   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0090, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.9943,
5.9943, 5.9943, 5.9943, 5.9943, 5.9943, 5.9943, 5.9943, 5.9943,
        5.9943, 5.4967, 5.0424, 4.6192, 4.2186, 3.8346, 3.4628, 3.1001, 2.7440,
        2.3929, 2.0454, 1.7008, 1.3582, 1.0172, 0.6773, 0.3383]) return=
122329.26544005591
probs of actions:  tensor([0.9989, 0.9987, 0.9990, 0.9990, 0.9989, 0.9987,
0.9991, 0.9988, 0.9990,
        0.9990, 0.9990, 0.9991, 0.9994, 0.9997, 0.9995, 0.9996, 0.9996, 0.9995,
        1.0000, 0.9996, 0.9998, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
```

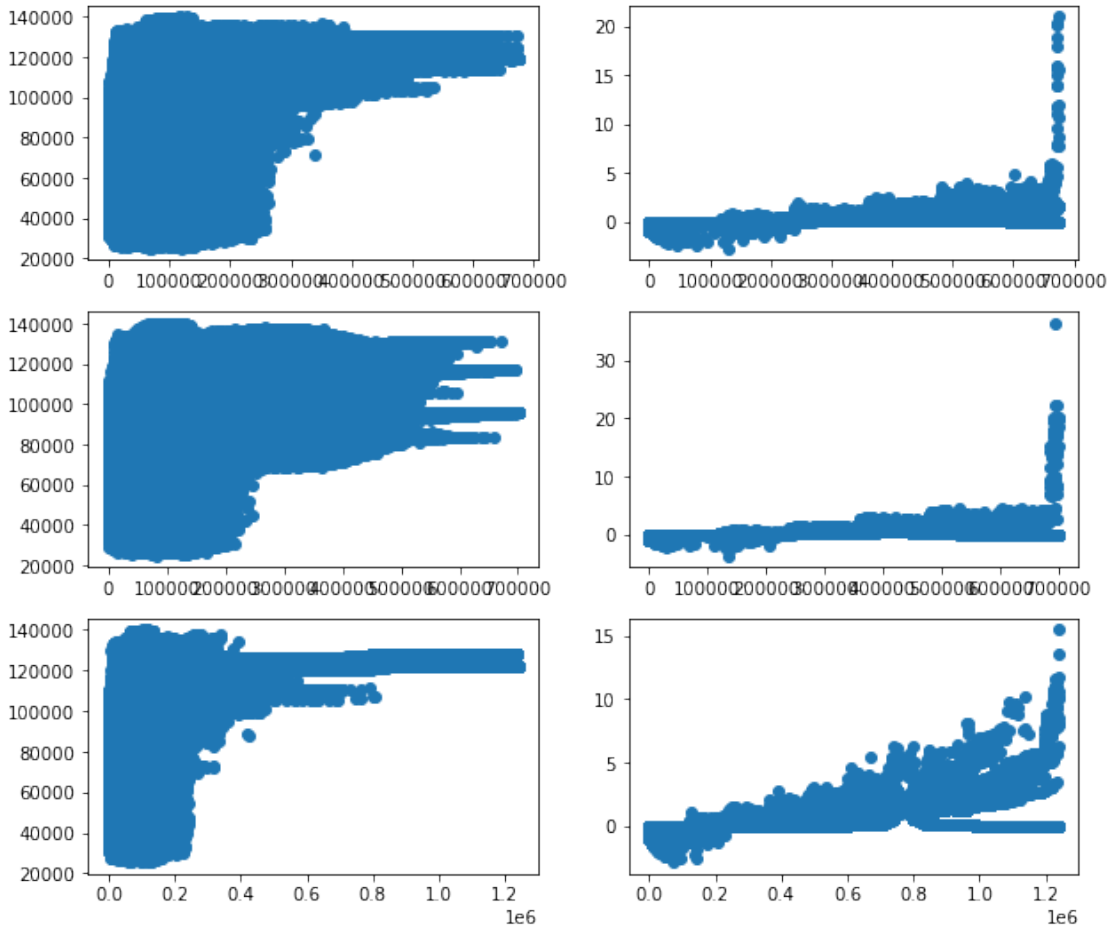
```
rewards:  tensor([0.4968, 0.5537, 0.5337, 0.5188, 0.5079, 0.4997, 0.4936,
0.4891, 0.4857,
        0.4832, 0.4813, 0.4798, 0.4788, 0.4780, 0.4774, 0.4769, 0.4766, 0.4764,
        0.4762, 0.4760, 0.4759, 0.4758, 0.4758, 0.4757, 0.4901])
finalReturns:  tensor([1.6596, 1.6740, 1.6471, 1.5904, 1.5122, 1.4182, 1.3126,
1.1985, 1.0779,
        0.9527, 0.8240, 0.6926, 0.5592, 0.4244, 0.2885, 0.1518])
----------------------------------------
iter  2  stage  8  ep  0   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0113, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.3414,
6.3414, 6.3414, 6.3414, 6.3414, 6.3414, 6.3414, 6.3414, 6.3414,
        5.8413, 5.3852, 4.9607, 4.5592, 4.1745, 3.8022, 3.4390, 3.0826, 2.7313,
        2.3837, 2.0390, 1.6963, 1.3552, 1.0153, 0.6763, 0.3379]) return=
122329.26544005591
probs of actions:  tensor([0.9989, 0.9987, 0.9990, 0.9990, 0.9989, 0.9987,
0.9991, 0.9988, 0.9990,
        0.9990, 0.9990, 0.9991, 0.9994, 0.9997, 0.9995, 0.9996, 0.9996, 0.9995,
        1.0000, 0.9996, 0.9998, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.5537, 0.5337, 0.5188, 0.5079, 0.4997, 0.4936,
0.4891, 0.4857,
        0.4832, 0.4813, 0.4798, 0.4788, 0.4780, 0.4774, 0.4769, 0.4766, 0.4764,
        0.4762, 0.4760, 0.4759, 0.4758, 0.4758, 0.4757, 0.4901])
finalReturns:  tensor([1.7982, 1.8126, 1.7855, 1.7288, 1.6505, 1.5564, 1.4507,
1.3365, 1.2159,
        1.0906, 0.9619, 0.8304, 0.6971, 0.5623, 0.4263, 0.2896, 0.1522])
----------------------------------------
iter  2  stage  7  ep  179   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0132, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.6914,
6.6914, 6.6914, 6.6914, 6.6914, 6.6914, 6.6914, 6.6914, 6.1879,
        5.7293, 5.3031, 4.9003, 4.5146, 4.1416, 3.7779, 3.4212, 3.0696, 2.7218,
        2.3769, 2.0341, 1.6929, 1.3529, 1.0139, 0.6754, 0.3375]) return=
122329.26544005591
probs of actions:  tensor([0.9989, 0.9987, 0.9990, 0.9990, 0.9989, 0.9987,
0.9991, 0.9990, 0.9991,
        0.9991, 0.9990, 0.9991, 0.9994, 0.9997, 0.9995, 0.9996, 0.9996, 0.9995,
        1.0000, 0.9996, 0.9998, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.5537, 0.5337, 0.5188, 0.5079, 0.4997, 0.4936,
0.4891, 0.4857,
        0.4832, 0.4813, 0.4798, 0.4788, 0.4780, 0.4774, 0.4769, 0.4766, 0.4764,
        0.4762, 0.4760, 0.4759, 0.4758, 0.4758, 0.4757, 0.4901])
```

```
finalReturns:  tensor([1.9373, 1.9517, 1.9246, 1.8677, 1.7892, 1.6950, 1.5892,
1.4749, 1.3543,
        1.2290, 1.1001, 0.9687, 0.8353, 0.7005, 0.5645, 0.4278, 0.2904, 0.1526])
----------------------------------------
iter  2  stage  6  ep  0   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0154, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.0453,
7.0453, 7.0453, 7.0453, 7.0453, 7.0453, 7.0453, 6.5372, 6.0754,
        5.6468, 5.2423, 4.8554, 4.4815, 4.1171, 3.7599, 3.4079, 3.0598, 2.7147,
        2.3717, 2.0305, 1.6904, 1.3512, 1.0128, 0.6748, 0.3373]) return=
122329.26544005591
probs of actions:  tensor([0.9989, 0.9987, 0.9990, 0.9990, 0.9989, 0.9987,
0.9991, 0.9990, 0.9991,
        0.9991, 0.9990, 0.9991, 0.9994, 0.9997, 0.9995, 0.9996, 0.9996, 0.9995,
        1.0000, 0.9996, 0.9998, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.5537, 0.5337, 0.5188, 0.5079, 0.4997, 0.4936,
0.4891, 0.4857,
        0.4832, 0.4813, 0.4798, 0.4788, 0.4780, 0.4774, 0.4769, 0.4766, 0.4764,
        0.4762, 0.4760, 0.4759, 0.4758, 0.4758, 0.4757, 0.4901])
finalReturns:  tensor([2.0770, 2.0914, 2.0642, 2.0071, 1.9284, 1.8340, 1.7281,
1.6137, 1.4930,
        1.3676, 1.2387, 1.1072, 0.9738, 0.8389, 0.7030, 0.5662, 0.4288, 0.2910,
        0.1528])
----------------------------------------
iter  2  stage  5  ep  179   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0167, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.4046,
7.4046, 7.4046, 7.4046, 7.4046, 7.4046, 6.8905, 6.4243, 5.9926,
        5.5858, 5.1973, 4.8221, 4.4568, 4.0989, 3.7464, 3.3979, 3.0525, 2.7094,
        2.3679, 2.0277, 1.6885, 1.3500, 1.0120, 0.6744, 0.3371]) return=
122329.26544005591
probs of actions:  tensor([0.9990, 0.9988, 0.9990, 0.9990, 0.9990, 0.9990,
0.9993, 0.9991, 0.9991,
        0.9992, 0.9991, 0.9992, 0.9994, 0.9998, 0.9995, 0.9996, 0.9997, 0.9995,
        1.0000, 0.9996, 0.9998, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.5537, 0.5337, 0.5188, 0.5079, 0.4997, 0.4936,
0.4891, 0.4857,
        0.4832, 0.4813, 0.4798, 0.4788, 0.4780, 0.4774, 0.4769, 0.4766, 0.4764,
        0.4762, 0.4760, 0.4759, 0.4758, 0.4758, 0.4757, 0.4901])
finalReturns:  tensor([2.2174, 2.2318, 2.2043, 2.1470, 2.0681, 1.9735, 1.8674,
1.7528, 1.6320,
        1.5065, 1.3775, 1.2460, 1.1126, 0.9777, 0.8417, 0.7049, 0.5675, 0.4296,
```

```
        0.2915, 0.1530])
---------------------------------------
iter  2  stage  4  ep  42   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0198, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.7714,
7.7714, 7.7714, 7.7714, 7.7714, 7.2491, 6.7771, 6.3412, 5.9313,
        5.5405, 5.1637, 4.7972, 4.4384, 4.0852, 3.7363, 3.3905, 3.0470, 2.7054,
        2.3650, 2.0257, 1.6871, 1.3490, 1.0114, 0.6740, 0.3369]) return=
122329.26544005591
probs of actions:  tensor([0.9990, 0.9988, 0.9990, 0.9990, 0.9990, 0.9990,
0.9993, 0.9991, 0.9991,
        0.9992, 0.9991, 0.9991, 0.9994, 0.9998, 0.9995, 0.9996, 0.9997, 0.9995,
        1.0000, 0.9996, 0.9998, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.5537, 0.5337, 0.5188, 0.5079, 0.4997, 0.4936,
0.4891, 0.4857,
        0.4832, 0.4813, 0.4798, 0.4788, 0.4780, 0.4774, 0.4769, 0.4766, 0.4764,
        0.4762, 0.4760, 0.4759, 0.4758, 0.4758, 0.4757, 0.4901])
finalReturns:  tensor([2.3585, 2.3729, 2.3452, 2.2875, 2.2082, 2.1134, 2.0070,
1.8922, 1.7713,
        1.6456, 1.5166, 1.3850, 1.2515, 1.1165, 0.9805, 0.8437, 0.7063, 0.5684,
        0.4303, 0.2918, 0.1532])
---------------------------------------
iter  2  stage  3  ep  0   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0227, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.1482,
8.1482, 8.1482, 8.1482, 7.6150, 7.1352, 6.6936, 6.2797, 5.8859,
        5.5069, 5.1387, 4.7787, 4.4246, 4.0750, 3.7287, 3.3849, 3.0429, 2.7024,
        2.3629, 2.0241, 1.6860, 1.3483, 1.0109, 0.6738, 0.3368]) return=
122329.26544005591
probs of actions:  tensor([0.9990, 0.9988, 0.9990, 0.9990, 0.9990, 0.9990,
0.9993, 0.9991, 0.9991,
        0.9992, 0.9991, 0.9991, 0.9994, 0.9998, 0.9995, 0.9996, 0.9997, 0.9995,
        1.0000, 0.9996, 0.9998, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.5537, 0.5337, 0.5188, 0.5079, 0.4997, 0.4936,
0.4891, 0.4857,
        0.4832, 0.4813, 0.4798, 0.4788, 0.4780, 0.4774, 0.4769, 0.4766, 0.4764,
        0.4762, 0.4760, 0.4759, 0.4758, 0.4758, 0.4757, 0.4901])
finalReturns:  tensor([2.5005, 2.5149, 2.4868, 2.4287, 2.3490, 2.2537, 2.1470,
2.0320, 1.9108,
        1.7850, 1.6559, 1.5242, 1.3906, 1.2556, 1.1195, 0.9827, 0.8452, 0.7074,
        0.5692, 0.4307, 0.2921, 0.1533])
---------------------------------------
```

```
iter  2  stage  2  ep  0   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0259, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.5388,
8.5388, 8.5388, 7.9907, 7.5004, 7.0513, 6.6318, 6.2340, 5.8521,
        5.4817, 5.1201, 4.7648, 4.4143, 4.0673, 3.7230, 3.3807, 3.0399, 2.7002,
        2.3613, 2.0230, 1.6852, 1.3478, 1.0106, 0.6736, 0.3367]) return=
122329.26544005591
probs of actions:  tensor([0.9990, 0.9988, 0.9990, 0.9990, 0.9990, 0.9990,
0.9993, 0.9991, 0.9991,
        0.9992, 0.9991, 0.9991, 0.9994, 0.9998, 0.9995, 0.9996, 0.9997, 0.9995,
        1.0000, 0.9996, 0.9998, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.5537, 0.5337, 0.5188, 0.5079, 0.4997, 0.4936,
0.4891, 0.4857,
        0.4832, 0.4813, 0.4798, 0.4788, 0.4780, 0.4774, 0.4769, 0.4766, 0.4764,
        0.4762, 0.4760, 0.4759, 0.4758, 0.4758, 0.4757, 0.4901])
finalReturns:  tensor([2.6436, 2.6580, 2.6294, 2.5707, 2.4905, 2.3946, 2.2875,
2.1722, 2.0507,
        1.9247, 1.7953, 1.6635, 1.5298, 1.3948, 1.2586, 1.1218, 0.9843, 0.8464,
        0.7082, 0.5697, 0.4310, 0.2923, 0.1534])
----------------------------------------
iter  2  stage  1  ep  130   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0258, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.9480,
8.9480, 8.3799, 7.8753, 7.4160, 6.9892, 6.5860, 6.2001, 5.8268,
        5.4630, 5.1061, 4.7544, 4.4066, 4.0616, 3.7188, 3.3776, 3.0376, 2.6985,
        2.3601, 2.0221, 1.6846, 1.3473, 1.0103, 0.6734, 0.3367]) return=
122329.26544005591
probs of actions:  tensor([0.9990, 0.9990, 0.9992, 0.9992, 0.9992, 0.9991,
0.9994, 0.9993, 0.9991,
        0.9993, 0.9991, 0.9992, 0.9995, 0.9998, 0.9995, 0.9996, 0.9997, 0.9996,
        1.0000, 0.9996, 0.9999, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.5537, 0.5337, 0.5188, 0.5079, 0.4997, 0.4936,
0.4891, 0.4857,
        0.4832, 0.4813, 0.4798, 0.4788, 0.4780, 0.4774, 0.4769, 0.4766, 0.4764,
        0.4762, 0.4760, 0.4759, 0.4758, 0.4758, 0.4757, 0.4901])
finalReturns:  tensor([2.7881, 2.8025, 2.7734, 2.7139, 2.6328, 2.5363, 2.4286,
2.3128, 2.1909,
        2.0646, 1.9351, 1.8031, 1.6692, 1.5341, 1.3979, 1.2609, 1.1234, 0.9855,
        0.8473, 0.7088, 0.5701, 0.4313, 0.2924, 0.1534])
----------------------------------------
iter  2  stage  0  ep  0   adversary:  AdversaryModes.fight_lb_132
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
```

```
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0291, grad_fn=<NegBackward0>)   ,  base rewards= tensor([9.2984,
8.7872, 8.2633, 7.7903, 7.3535, 6.9431, 6.5519, 6.1747, 5.8080,
        5.4490, 5.0957, 4.7466, 4.4008, 4.0573, 3.7156, 3.3752, 3.0359, 2.6972,
        2.3591, 2.0215, 1.6842, 1.3470, 1.0101, 0.6733, 0.3366]) return=
122329.26544005591
probs of actions:  tensor([0.9990, 0.9990, 0.9992, 0.9992, 0.9992, 0.9991,
0.9994, 0.9993, 0.9991,
        0.9993, 0.9991, 0.9992, 0.9995, 0.9998, 0.9995, 0.9996, 0.9997, 0.9996,
        1.0000, 0.9996, 0.9999, 1.0000, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.5537, 0.5337, 0.5188, 0.5079, 0.4997, 0.4936,
0.4891, 0.4857,
        0.4832, 0.4813, 0.4798, 0.4788, 0.4780, 0.4774, 0.4769, 0.4766, 0.4764,
        0.4762, 0.4760, 0.4759, 0.4758, 0.4758, 0.4757, 0.4901])
finalReturns:  tensor([2.9345, 2.9489, 2.9190, 2.8584, 2.7763, 2.6789, 2.5704,
2.4539, 2.3316,
        2.2049, 2.0751, 1.9428, 1.8088, 1.6735, 1.5372, 1.4002, 1.2627, 1.1247,
        0.9864, 0.8479, 0.7092, 0.5704, 0.4315, 0.2925, 0.1535])
0,[1e-05,1][1, 10000, 1, 1],1683049303 saved
[1239451, 'tensor([0., 0., 0., 0., 0., 0., 1., 0., 0., 0., 0., 0.])',
122329.26544005591, 95669.58206529099, 0.029126999899744987, 1e-05, 1, 0,
'tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12,\n        12, 12, 12, 12, 12, 12,  0])', '[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.\n 1.]', '0,[1e-05,1][1, 10000, 1,
1],1683049303', 25, 50, 157611.33342506486, 175538.2609849781,
68694.09895647192, 135313.23466666666, 132439.05866666668, 122329.26544005591,
122329.26544005591, 130254.64915679736, 130254.64915679736, 80045.82189636558,
122329.26544005591, 130254.64915679736]
```




```
policy reset
----------------------------------------
iter  0  stage  24  ep  99999   adversary:  AdversaryModes.fight_125
  actions:  tensor([0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 4, 0, 0,
0, 1, 0, 0,
        0])
loss=  tensor(-0., grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.4671,
0.4671, 0.4671, 0.4671, 0.4671, 0.4671, 0.4671, 0.4671, 0.4671,
        0.4671, 0.4671, 0.4671, 0.4671, 0.4671, 0.4671, 0.4671, 0.4671, 0.4671,
        0.4671, 0.4671, 0.4671, 0.4671, 0.4671, 0.4671, 0.4671]) return=
118465.88424035063
probs of actions:  tensor([0.9225, 0.9092, 0.0133, 0.9142, 0.9103, 0.9245,
0.9002, 0.9208, 0.9088,
        0.9158, 0.9157, 0.9228, 0.9095, 0.8977, 0.0451, 0.9123, 0.9046, 0.0062,
        0.9008, 0.9034, 0.9077, 0.0532, 0.9217, 0.8866, 0.9858],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5112, 0.4988, 0.4892, 0.4897, 0.4828, 0.4776, 0.4738,
0.4709, 0.4688,
```

```
        0.4672, 0.4660, 0.4651, 0.4644, 0.4639, 0.4634, 0.4667, 0.4656, 0.4632,
        0.4779, 0.4740, 0.4711, 0.4688, 0.4707, 0.4686, 0.4671])
finalReturns:  tensor([0.])
----------------------------------------
iter  0  stage  23  ep  99999   adversary:  AdversaryModes.fight_125
  actions:  tensor([ 6,  0,  0,  9,  0,  0,  0,  9,  7,  0,  0, 12,  1, 13,  0,
0,  7,  8,
         0,  9,  0,  7,  9,  9,  0])
loss=  tensor(0.0365, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.0671,
1.0671, 1.0671, 1.0671, 1.0671, 1.0671, 1.0671, 1.0671, 1.0671,
        1.0671, 1.0671, 1.0671, 1.0671, 1.0671, 1.0671, 1.0671, 1.0671, 1.0671,
        1.0671, 1.0671, 1.0671, 1.0671, 1.0671, 1.0671, 0.5230]) return=
128711.60607356537
probs of actions:  tensor([0.0279, 0.3195, 0.2609, 0.3040, 0.4547, 0.4141,
0.3812, 0.2469, 0.0893,
        0.2782, 0.3144, 0.0407, 0.0230, 0.0346, 0.2974, 0.2232, 0.0981, 0.1333,
        0.3403, 0.2705, 0.3291, 0.1090, 0.2630, 0.2326, 0.9983],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5076, 0.5202, 0.5054, 0.4864, 0.5183, 0.5040, 0.4934,
0.4775, 0.5065,
        0.5240, 0.5082, 0.4821, 0.5306, 0.4999, 0.5501, 0.5274, 0.5059, 0.5171,
        0.5368, 0.5095, 0.5360, 0.5122, 0.5202, 0.5360, 0.5561])
finalReturns:  tensor([0.0250, 0.0331])
----------------------------------------
iter  0  stage  22  ep  99999   adversary:  AdversaryModes.fight_125
  actions:  tensor([17, 21, 21, 21, 21, 21, 14, 21,  7, 11, 21, 21, 13,  9, 21,
21, 13, 21,
        21, 21, 21, 21, 21, 17,  0])
loss=  tensor(0.4933, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.9320,
0.9320, 0.9320, 0.9320, 0.9320, 0.9320, 0.9320, 0.9320, 0.9320,
        0.9320, 0.9320, 0.9320, 0.9320, 0.9320, 0.9320, 0.9320, 0.9320, 0.9320,
        0.9320, 0.9320, 0.9320, 0.9320, 0.9320, 0.5697, 0.2649]) return=
90665.36375873255
probs of actions:  tensor([0.0206, 0.6746, 0.6706, 0.7113, 0.6599, 0.7061,
0.0543, 0.6362, 0.0054,
        0.0176, 0.6451, 0.6355, 0.1559, 0.0499, 0.6243, 0.6552, 0.1977, 0.5576,
        0.6014, 0.5997, 0.6143, 0.6144, 0.9103, 0.0191, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4823, 0.5165, 0.5707, 0.5052, 0.4586, 0.4250, 0.4250,
0.3600, 0.3919,
        0.3367, 0.2820, 0.2939, 0.3302, 0.3225, 0.2632, 0.2796, 0.3193, 0.2785,
        0.2913, 0.3010, 0.3084, 0.3140, 0.3182, 0.3366, 0.3559])
finalReturns:  tensor([0.0787, 0.1228, 0.0910])
----------------------------------------
iter  0  stage  21  ep  99999   adversary:  AdversaryModes.fight_125
  actions:  tensor([21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21,
21, 21, 21,
        21, 21, 21, 21, 21, 13,  0])
```

```
loss=  tensor(0.4876, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.5149,
1.5149, 1.5149, 1.5149, 1.5149, 1.5149, 1.5149, 1.5149, 1.5149,
        1.5149, 1.5149, 1.5149, 1.5149, 1.5149, 1.5149, 1.5149, 1.5149, 1.5149,
        1.5149, 1.5149, 1.5149, 1.5149, 1.0522, 0.6582, 0.3121]) return=
109929.65918662024
probs of actions:  tensor([0.9783, 0.9750, 0.9743, 0.9797, 0.9723, 0.9814,
0.9655, 0.9731, 0.9711,
        0.9687, 0.9743, 0.9726, 0.9711, 0.9715, 0.9721, 0.9763, 0.9660, 0.9519,
        0.9693, 0.9662, 0.9683, 0.9910, 0.9983, 0.1137, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4671, 0.5316, 0.5021, 0.4805, 0.4646, 0.4528, 0.4441,
0.4376, 0.4327,
        0.4291, 0.4264, 0.4244, 0.4228, 0.4217, 0.4208, 0.4202, 0.4197, 0.4194,
        0.4191, 0.4189, 0.4188, 0.4186, 0.4186, 0.4457, 0.4357])
finalReturns:  tensor([0.2037, 0.2478, 0.2232, 0.1236])
----------------------------------------
iter  0  stage  20  ep  99999   adversary:  AdversaryModes.fight_125
  actions:  tensor([21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21,
21, 21, 21,
        21, 21, 21, 21, 21, 21,  0])
loss=  tensor(0.0092, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.8031,
1.8031, 1.8031, 1.8031, 1.8031, 1.8031, 1.8031, 1.8031, 1.8031,
        1.8031, 1.8031, 1.8031, 1.8031, 1.8031, 1.8031, 1.8031, 1.8031, 1.8031,
        1.8031, 1.8031, 1.8031, 1.3403, 0.9462, 0.6001, 0.2879]) return=
109925.7013290539
probs of actions:  tensor([0.9984, 0.9981, 0.9979, 0.9986, 0.9977, 0.9987,
0.9970, 0.9980, 0.9974,
        0.9973, 0.9981, 0.9979, 0.9977, 0.9978, 0.9978, 0.9981, 0.9972, 0.9956,
        0.9976, 0.9975, 0.9985, 0.9999, 1.0000, 0.9697, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4671, 0.5316, 0.5021, 0.4805, 0.4646, 0.4528, 0.4441,
0.4376, 0.4327,
        0.4291, 0.4264, 0.4244, 0.4228, 0.4217, 0.4208, 0.4202, 0.4197, 0.4194,
        0.4191, 0.4189, 0.4188, 0.4186, 0.4186, 0.4185, 0.4625])
finalReturns:  tensor([0.3339, 0.3780, 0.3534, 0.2810, 0.1747])
----------------------------------------
iter  0  stage  19  ep  956   adversary:  AdversaryModes.fight_125
  actions:  tensor([21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21,
21, 21, 21,
        13, 21, 21, 21, 21, 21,  0])
loss=  tensor(0.0114, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.9943,
1.9943, 1.9943, 1.9943, 1.9943, 1.9943, 1.9943, 1.9943, 1.9943,
        1.9943, 1.9943, 1.9943, 1.9943, 1.9943, 1.9943, 1.9943, 1.9943, 1.9943,
        1.9943, 1.9943, 1.5581, 1.1825, 0.8495, 0.5466, 0.2654]) return=
109311.80214161768
probs of actions:  tensor([0.9987, 0.9985, 0.9984, 0.9989, 0.9982, 0.9989,
0.9977, 0.9984, 0.9980,
        0.9978, 0.9985, 0.9985, 0.9982, 0.9983, 0.9983, 0.9985, 0.9978, 0.9966,
```

```
        0.0010, 0.9990, 0.9989, 1.0000, 1.0000, 0.9682, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4671, 0.5316, 0.5021, 0.4805, 0.4646, 0.4528, 0.4441,
0.4376, 0.4327,
        0.4291, 0.4264, 0.4244, 0.4228, 0.4217, 0.4208, 0.4202, 0.4197, 0.4194,
        0.4463, 0.3921, 0.3986, 0.4035, 0.4071, 0.4099, 0.4561])
finalReturns:  tensor([0.4730, 0.5171, 0.4941, 0.4237, 0.3194, 0.1907])
----------------------------------------
iter  0  stage  18  ep  722   adversary:  AdversaryModes.fight_125
  actions:  tensor([21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21,
21, 21, 21,
        21, 21, 21, 21, 21, 21,  0])
loss=  tensor(0.0127, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.3320,
2.3320, 2.3320, 2.3320, 2.3320, 2.3320, 2.3320, 2.3320, 2.3320,
        2.3320, 2.3320, 2.3320, 2.3320, 2.3320, 2.3320, 2.3320, 2.3320, 2.3320,
        2.3320, 1.8688, 1.4745, 1.1282, 0.8159, 0.5279, 0.2575]) return=
109925.7013290539
probs of actions:  tensor([0.9988, 0.9987, 0.9985, 0.9990, 0.9984, 0.9990,
0.9979, 0.9986, 0.9981,
        0.9980, 0.9986, 0.9986, 0.9984, 0.9984, 0.9984, 0.9987, 0.9980, 0.9969,
        0.9990, 0.9993, 0.9990, 1.0000, 1.0000, 0.9694, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4671, 0.5316, 0.5021, 0.4805, 0.4646, 0.4528, 0.4441,
0.4376, 0.4327,
        0.4291, 0.4264, 0.4244, 0.4228, 0.4217, 0.4208, 0.4202, 0.4197, 0.4194,
        0.4191, 0.4189, 0.4188, 0.4186, 0.4186, 0.4185, 0.4625])
finalReturns:  tensor([0.6430, 0.6871, 0.6625, 0.5900, 0.4837, 0.3531, 0.2050])
----------------------------------------
iter  0  stage  17  ep  50770   adversary:  AdversaryModes.fight_125
  actions:  tensor([21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21,
21, 21, 21,
        21, 21, 21, 21, 21, 21,  0])
loss=  tensor(0.0102, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.5810,
2.5810, 2.5810, 2.5810, 2.5810, 2.5810, 2.5810, 2.5810, 2.5810,
        2.5810, 2.5810, 2.5810, 2.5810, 2.5810, 2.5810, 2.5810, 2.5810, 2.5810,
        2.1175, 1.7230, 1.3766, 1.0642, 0.7761, 0.5057, 0.2481]) return=
109925.7013290539
probs of actions:  tensor([0.9994, 0.9993, 0.9992, 0.9995, 0.9991, 0.9995,
0.9989, 0.9993, 0.9990,
        0.9989, 0.9993, 0.9993, 0.9992, 0.9992, 0.9992, 0.9993, 0.9989, 0.9990,
        0.9997, 1.0000, 0.9999, 1.0000, 1.0000, 0.9761, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4671, 0.5316, 0.5021, 0.4805, 0.4646, 0.4528, 0.4441,
0.4376, 0.4327,
        0.4291, 0.4264, 0.4244, 0.4228, 0.4217, 0.4208, 0.4202, 0.4197, 0.4194,
        0.4191, 0.4189, 0.4188, 0.4186, 0.4186, 0.4185, 0.4625])
finalReturns:  tensor([0.8134, 0.8575, 0.8329, 0.7604, 0.6541, 0.5235, 0.3753,
0.2144])
```

```
---------------------------------------
iter  0  stage  16  ep  39   adversary:  AdversaryModes.fight_125
  actions:  tensor([21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21,
21, 21, 21,
        21, 21, 21, 21, 21, 21,  0])
loss=  tensor(0.0118, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.8233,
2.8233, 2.8233, 2.8233, 2.8233, 2.8233, 2.8233, 2.8233, 2.8233,
        2.8233, 2.8233, 2.8233, 2.8233, 2.8233, 2.8233, 2.8233, 2.8233, 2.3595,
        1.9647, 1.6181, 1.3056, 1.0175, 0.7470, 0.4894, 0.2412]) return=
109925.7013290539
probs of actions:  tensor([0.9994, 0.9993, 0.9992, 0.9995, 0.9991, 0.9995,
0.9989, 0.9993, 0.9990,
        0.9989, 0.9993, 0.9993, 0.9992, 0.9992, 0.9992, 0.9993, 0.9990, 0.9990,
        0.9997, 1.0000, 0.9999, 1.0000, 1.0000, 0.9763, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4671, 0.5316, 0.5021, 0.4805, 0.4646, 0.4528, 0.4441,
0.4376, 0.4327,
        0.4291, 0.4264, 0.4244, 0.4228, 0.4217, 0.4208, 0.4202, 0.4197, 0.4194,
        0.4191, 0.4189, 0.4188, 0.4186, 0.4186, 0.4185, 0.4625])
finalReturns:  tensor([0.9908, 1.0349, 1.0103, 0.9377, 0.8314, 0.7007, 0.5526,
0.3917, 0.2213])
---------------------------------------
iter  0  stage  15  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21,
21, 21, 21,
        21, 21, 21, 21, 21, 21,  0])
loss=  tensor(0.0134, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.0609,
3.0609, 3.0609, 3.0609, 3.0609, 3.0609, 3.0609, 3.0609, 3.0609,
        3.0609, 3.0609, 3.0609, 3.0609, 3.0609, 3.0609, 3.0609, 2.5966, 2.2016,
        1.8547, 1.5420, 1.2538, 0.9832, 0.7255, 0.4773, 0.2361]) return=
109925.7013290539
probs of actions:  tensor([0.9994, 0.9993, 0.9992, 0.9995, 0.9991, 0.9995,
0.9989, 0.9993, 0.9990,
        0.9989, 0.9993, 0.9993, 0.9992, 0.9992, 0.9992, 0.9993, 0.9990, 0.9990,
        0.9997, 1.0000, 0.9999, 1.0000, 1.0000, 0.9763, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4671, 0.5316, 0.5021, 0.4805, 0.4646, 0.4528, 0.4441,
0.4376, 0.4327,
        0.4291, 0.4264, 0.4244, 0.4228, 0.4217, 0.4208, 0.4202, 0.4197, 0.4194,
        0.4191, 0.4189, 0.4188, 0.4186, 0.4186, 0.4185, 0.4625])
finalReturns:  tensor([1.1734, 1.2175, 1.1928, 1.1203, 1.0138, 0.8832, 0.7350,
0.5741, 0.4037,
        0.2265])
---------------------------------------
iter  0  stage  14  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21,
21, 21, 21,
        21, 21, 21, 21, 21, 21,  0])
```

```
loss=  tensor(0.0151, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.2953,
3.2953, 3.2953, 3.2953, 3.2953, 3.2953, 3.2953, 3.2953, 3.2953,
        3.2953, 3.2953, 3.2953, 3.2953, 3.2953, 3.2953, 2.8304, 2.4348, 2.0877,
        1.7748, 1.4864, 1.2157, 0.9579, 0.7097, 0.4684, 0.2323]) return=
109925.7013290539
probs of actions:  tensor([0.9994, 0.9993, 0.9992, 0.9995, 0.9991, 0.9995,
0.9989, 0.9993, 0.9990,
        0.9989, 0.9993, 0.9993, 0.9992, 0.9992, 0.9992, 0.9993, 0.9990, 0.9990,
        0.9997, 1.0000, 0.9999, 1.0000, 1.0000, 0.9763, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4671, 0.5316, 0.5021, 0.4805, 0.4646, 0.4528, 0.4441,
0.4376, 0.4327,
        0.4291, 0.4264, 0.4244, 0.4228, 0.4217, 0.4208, 0.4202, 0.4197, 0.4194,
        0.4191, 0.4189, 0.4188, 0.4186, 0.4186, 0.4185, 0.4625])
finalReturns:  tensor([1.3598, 1.4039, 1.3793, 1.3067, 1.2002, 1.0695, 0.9213,
0.7603, 0.5899,
        0.4127, 0.2303])
---------------------------------------
iter  0  stage  13  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21,
21, 21, 21,
        21, 21, 21, 21, 21, 21,  0])
loss=  tensor(0.0171, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.5276,
3.5276, 3.5276, 3.5276, 3.5276, 3.5276, 3.5276, 3.5276, 3.5276,
        3.5276, 3.5276, 3.5276, 3.5276, 3.5276, 3.0618, 2.6657, 2.3181, 2.0049,
        1.7163, 1.4455, 1.1876, 0.9392, 0.6979, 0.4617, 0.2294]) return=
109925.7013290539
probs of actions:  tensor([0.9995, 0.9993, 0.9992, 0.9995, 0.9991, 0.9995,
0.9989, 0.9993, 0.9990,
        0.9989, 0.9993, 0.9993, 0.9992, 0.9992, 0.9992, 0.9993, 0.9990, 0.9990,
        0.9997, 1.0000, 0.9999, 1.0000, 1.0000, 0.9763, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4671, 0.5316, 0.5021, 0.4805, 0.4646, 0.4528, 0.4441,
0.4376, 0.4327,
        0.4291, 0.4264, 0.4244, 0.4228, 0.4217, 0.4208, 0.4202, 0.4197, 0.4194,
        0.4191, 0.4189, 0.4188, 0.4186, 0.4186, 0.4185, 0.4625])
finalReturns:  tensor([1.5493, 1.5934, 1.5686, 1.4960, 1.3894, 1.2587, 1.1104,
0.9494, 0.7790,
        0.6017, 0.4193, 0.2331])
---------------------------------------
iter  0  stage  12  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21,
21, 21, 21,
        21, 21, 21, 21, 21, 21,  0])
loss=  tensor(0.0192, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.7587,
3.7587, 3.7587, 3.7587, 3.7587, 3.7587, 3.7587, 3.7587, 3.7587,
        3.7587, 3.7587, 3.7587, 3.7587, 3.2918, 2.8949, 2.5468, 2.2332, 1.9443,
        1.6732, 1.4152, 1.1667, 0.9253, 0.6891, 0.4567, 0.2273]) return=
```

```
109925.7013290539
probs of actions:  tensor([0.9995, 0.9993, 0.9992, 0.9995, 0.9991, 0.9995,
0.9989, 0.9993, 0.9990,
        0.9989, 0.9993, 0.9993, 0.9992, 0.9992, 0.9992, 0.9993, 0.9990, 0.9990,
        0.9997, 1.0000, 0.9999, 1.0000, 1.0000, 0.9763, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4671, 0.5316, 0.5021, 0.4805, 0.4646, 0.4528, 0.4441,
0.4376, 0.4327,
        0.4291, 0.4264, 0.4244, 0.4228, 0.4217, 0.4208, 0.4202, 0.4197, 0.4194,
        0.4191, 0.4189, 0.4188, 0.4186, 0.4186, 0.4185, 0.4625])
finalReturns:  tensor([1.7410, 1.7851, 1.7603, 1.6875, 1.5809, 1.4501, 1.3018,
1.1407, 0.9703,
        0.7929, 0.6105, 0.4243, 0.2352])
----------------------------------------
iter  0  stage  11  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21,
21, 21, 21,
        21, 21, 21, 21, 21, 21,  0])
loss=  tensor(0.0211, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.9896,
3.9896, 3.9896, 3.9896, 3.9896, 3.9896, 3.9896, 3.9896, 3.9896,
        3.9896, 3.9896, 3.9896, 3.5211, 3.1232, 2.7744, 2.4602, 2.1710, 1.8996,
        1.6414, 1.3928, 1.1512, 0.9149, 0.6825, 0.4530, 0.2257]) return=
109925.7013290539
probs of actions:  tensor([0.9995, 0.9993, 0.9992, 0.9995, 0.9991, 0.9995,
0.9989, 0.9993, 0.9990,
        0.9989, 0.9993, 0.9993, 0.9992, 0.9992, 0.9992, 0.9993, 0.9990, 0.9990,
        0.9997, 1.0000, 0.9999, 1.0000, 1.0000, 0.9763, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4671, 0.5316, 0.5021, 0.4805, 0.4646, 0.4528, 0.4441,
0.4376, 0.4327,
        0.4291, 0.4264, 0.4244, 0.4228, 0.4217, 0.4208, 0.4202, 0.4197, 0.4194,
        0.4191, 0.4189, 0.4188, 0.4186, 0.4186, 0.4185, 0.4625])
finalReturns:  tensor([1.9345, 1.9786, 1.9537, 1.8808, 1.7741, 1.6431, 1.4947,
1.3336, 1.1631,
        0.9857, 0.8033, 0.6170, 0.4280, 0.2368])
----------------------------------------
iter  0  stage  10  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21,
21, 21, 21,
        21, 21, 21, 21, 21, 21,  0])
loss=  tensor(0.0233, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.2210,
4.2210, 4.2210, 4.2210, 4.2210, 4.2210, 4.2210, 4.2210, 4.2210,
        4.2210, 4.2210, 3.7505, 3.3512, 3.0014, 2.6866, 2.3968, 2.1251, 1.8666,
        1.6178, 1.3761, 1.1397, 0.9072, 0.6777, 0.4503, 0.2245]) return=
109925.7013290539
probs of actions:  tensor([0.9995, 0.9993, 0.9992, 0.9995, 0.9991, 0.9995,
0.9989, 0.9993, 0.9990,
        0.9989, 0.9993, 0.9993, 0.9992, 0.9992, 0.9992, 0.9993, 0.9990, 0.9990,
```

```
        0.9997, 1.0000, 0.9999, 1.0000, 1.0000, 0.9763, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4671, 0.5316, 0.5021, 0.4805, 0.4646, 0.4528, 0.4441,
0.4376, 0.4327,
        0.4291, 0.4264, 0.4244, 0.4228, 0.4217, 0.4208, 0.4202, 0.4197, 0.4194,
        0.4191, 0.4189, 0.4188, 0.4186, 0.4186, 0.4185, 0.4625])
finalReturns:  tensor([2.1294, 2.1735, 2.1485, 2.0755, 1.9686, 1.8375, 1.6890,
1.5278, 1.3572,
        1.1798, 0.9973, 0.8110, 0.6219, 0.4308, 0.2380])
---------------------------------------
iter  0  stage  9  ep  40   adversary:  AdversaryModes.fight_125
  actions:  tensor([21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21,
21, 21, 21,
        21, 21, 21, 21, 21, 21,  0])
loss=  tensor(0.0255, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.4539,
4.4539, 4.4539, 4.4539, 4.4539, 4.4539, 4.4539, 4.4539, 4.4539,
        4.4539, 3.9807, 3.5795, 3.2284, 2.9127, 2.6222, 2.3500, 2.0912, 1.8421,
        1.6002, 1.3637, 1.1311, 0.9014, 0.6740, 0.4482, 0.2236]) return=
109925.7013290539
probs of actions:  tensor([0.9994, 0.9993, 0.9992, 0.9995, 0.9991, 0.9995,
0.9989, 0.9993, 0.9990,
        0.9990, 0.9994, 0.9994, 0.9992, 0.9993, 0.9993, 0.9993, 0.9990, 0.9990,
        0.9997, 1.0000, 0.9999, 1.0000, 1.0000, 0.9760, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4671, 0.5316, 0.5021, 0.4805, 0.4646, 0.4528, 0.4441,
0.4376, 0.4327,
        0.4291, 0.4264, 0.4244, 0.4228, 0.4217, 0.4208, 0.4202, 0.4197, 0.4194,
        0.4191, 0.4189, 0.4188, 0.4186, 0.4186, 0.4185, 0.4625])
finalReturns:  tensor([2.3256, 2.3697, 2.3445, 2.2713, 2.1642, 2.0330, 1.8843,
1.7229, 1.5523,
        1.3748, 1.1922, 1.0059, 0.8168, 0.6256, 0.4328, 0.2389])
---------------------------------------
iter  0  stage  8  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21,
21, 21, 21,
        21, 21, 21, 21, 21, 21,  0])
loss=  tensor(0.0287, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.6893,
4.6893, 4.6893, 4.6893, 4.6893, 4.6893, 4.6893, 4.6893, 4.6893,
        4.2125, 3.8088, 3.4559, 3.1389, 2.8476, 2.5747, 2.3154, 2.0660, 1.8238,
        1.5871, 1.3544, 1.1246, 0.8971, 0.6713, 0.4467, 0.2230]) return=
109925.7013290539
probs of actions:  tensor([0.9994, 0.9993, 0.9992, 0.9995, 0.9991, 0.9995,
0.9989, 0.9993, 0.9990,
        0.9990, 0.9994, 0.9994, 0.9992, 0.9993, 0.9993, 0.9993, 0.9990, 0.9990,
        0.9997, 1.0000, 0.9999, 1.0000, 1.0000, 0.9759, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4671, 0.5316, 0.5021, 0.4805, 0.4646, 0.4528, 0.4441,
0.4376, 0.4327,
```

```
        0.4291, 0.4264, 0.4244, 0.4228, 0.4217, 0.4208, 0.4202, 0.4197, 0.4194,
        0.4191, 0.4189, 0.4188, 0.4186, 0.4186, 0.4185, 0.4625])
finalReturns:  tensor([2.5229, 2.5670, 2.5416, 2.4681, 2.3608, 2.2293, 2.0804,
1.9189, 1.7481,
        1.5705, 1.3879, 1.2015, 1.0123, 0.8211, 0.6283, 0.4344, 0.2396])
----------------------------------------
iter  0  stage  7  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21,
21, 21, 21,
        21, 21, 21, 21, 21, 21,  0])
loss=  tensor(0.0314, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.9285,
4.9285, 4.9285, 4.9285, 4.9285, 4.9285, 4.9285, 4.9285, 4.4468,
        4.0397, 3.6845, 3.3658, 3.0733, 2.7996, 2.5396, 2.2897, 2.0472, 1.8102,
        1.5773, 1.3474, 1.1198, 0.8939, 0.6692, 0.4455, 0.2225]) return=
109925.7013290539
probs of actions:  tensor([0.9994, 0.9993, 0.9992, 0.9995, 0.9991, 0.9995,
0.9989, 0.9993, 0.9990,
        0.9990, 0.9994, 0.9994, 0.9993, 0.9993, 0.9993, 0.9993, 0.9990, 0.9990,
        0.9997, 1.0000, 0.9999, 1.0000, 1.0000, 0.9759, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4671, 0.5316, 0.5021, 0.4805, 0.4646, 0.4528, 0.4441,
0.4376, 0.4327,
        0.4291, 0.4264, 0.4244, 0.4228, 0.4217, 0.4208, 0.4202, 0.4197, 0.4194,
        0.4191, 0.4189, 0.4188, 0.4186, 0.4186, 0.4185, 0.4625])
finalReturns:  tensor([2.7213, 2.7654, 2.7398, 2.6659, 2.5582, 2.4264, 2.2773,
2.1155, 1.9446,
        1.7669, 1.5841, 1.3977, 1.2084, 1.0172, 0.8243, 0.6304, 0.4355, 0.2401])
----------------------------------------
iter  0  stage  6  ep  40   adversary:  AdversaryModes.fight_125
  actions:  tensor([21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21,
21, 21, 21,
        21, 21, 21, 21, 21, 21,  0])
loss=  tensor(0.0331, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.1729,
5.1729, 5.1729, 5.1729, 5.1729, 5.1729, 5.1729, 4.6847, 4.2732,
        3.9148, 3.5939, 3.2997, 3.0249, 2.7641, 2.5135, 2.2706, 2.0333, 1.8001,
        1.5700, 1.3423, 1.1162, 0.8915, 0.6677, 0.4446, 0.2221]) return=
109925.7013290539
probs of actions:  tensor([0.9995, 0.9993, 0.9992, 0.9996, 0.9991, 0.9995,
0.9990, 0.9993, 0.9991,
        0.9991, 0.9994, 0.9994, 0.9993, 0.9993, 0.9993, 0.9994, 0.9991, 0.9990,
        0.9997, 1.0000, 0.9999, 1.0000, 1.0000, 0.9763, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4671, 0.5316, 0.5021, 0.4805, 0.4646, 0.4528, 0.4441,
0.4376, 0.4327,
        0.4291, 0.4264, 0.4244, 0.4228, 0.4217, 0.4208, 0.4202, 0.4197, 0.4194,
        0.4191, 0.4189, 0.4188, 0.4186, 0.4186, 0.4185, 0.4625])
finalReturns:  tensor([2.9210, 2.9651, 2.9390, 2.8647, 2.7565, 2.6243, 2.4748,
2.3128, 2.1416,
```

```
        1.9637, 1.7808, 1.5943, 1.4050, 1.2136, 1.0208, 0.8268, 0.6319, 0.4364,
        0.2404])
----------------------------------------
iter  0  stage  5  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21,
21, 13, 21,
        21, 21, 21, 21, 21, 21,  0])
loss=  tensor(12.9543, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([5.4246, 5.4246, 5.4246, 5.4246, 5.4246, 5.4246, 4.9277, 4.5102, 4.1476,
        3.8237, 3.5274, 3.2510, 2.9890, 2.7377, 2.4941, 2.2563, 2.0228, 1.7925,
        1.5646, 1.3384, 1.1135, 0.8897, 0.6665, 0.4440, 0.2218]) return=
109226.96100827814
probs of actions:  tensor([9.9946e-01, 9.9933e-01, 9.9919e-01, 9.9955e-01,
9.9913e-01, 9.9952e-01,
        9.9900e-01, 9.9935e-01, 9.9914e-01, 9.9909e-01, 9.9942e-01, 9.9944e-01,
        9.9932e-01, 9.9934e-01, 9.9934e-01, 9.9939e-01, 5.4956e-04, 9.9907e-01,
        9.9969e-01, 1.0000e+00, 9.9992e-01, 1.0000e+00, 1.0000e+00, 9.7597e-01,
        1.0000e+00], grad_fn=<ExpBackward0>)
rewards:  tensor([0.4671, 0.5316, 0.5021, 0.4805, 0.4646, 0.4528, 0.4441,
0.4376, 0.4327,
        0.4291, 0.4264, 0.4244, 0.4228, 0.4217, 0.4208, 0.4202, 0.4469, 0.3925,
        0.3989, 0.4037, 0.4073, 0.4101, 0.4121, 0.4137, 0.4589])
finalReturns:  tensor([3.0521, 3.0962, 3.0697, 2.9947, 2.8859, 2.7531, 2.6032,
2.4408, 2.2693,
        2.0912, 1.9081, 1.7214, 1.5048, 1.3402, 1.1675, 0.9886, 0.8051, 0.6182,
        0.4286, 0.2371])
----------------------------------------
iter  0  stage  4  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21,
21, 21, 21,
        21, 21, 21, 21, 21, 21,  0])
loss=  tensor(0.0390, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.6865,
5.6865, 5.6865, 5.6865, 5.6865, 5.1778, 4.7521, 4.3839, 4.0560,
        3.7569, 3.4784, 3.2149, 2.9624, 2.7181, 2.4797, 2.2457, 2.0150, 1.7868,
        1.5605, 1.3355, 1.1115, 0.8883, 0.6657, 0.4435, 0.2216]) return=
109925.7013290539
probs of actions:  tensor([0.9995, 0.9993, 0.9992, 0.9996, 0.9991, 0.9995,
0.9990, 0.9993, 0.9991,
        0.9991, 0.9994, 0.9994, 0.9993, 0.9993, 0.9993, 0.9994, 0.9990, 0.9990,
        0.9997, 1.0000, 0.9999, 1.0000, 1.0000, 0.9762, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4671, 0.5316, 0.5021, 0.4805, 0.4646, 0.4528, 0.4441,
0.4376, 0.4327,
        0.4291, 0.4264, 0.4244, 0.4228, 0.4217, 0.4208, 0.4202, 0.4197, 0.4194,
        0.4191, 0.4189, 0.4188, 0.4186, 0.4186, 0.4185, 0.4625])
finalReturns:  tensor([3.3248, 3.3689, 3.3417, 3.2659, 3.1562, 3.0226, 2.8720,
2.7091, 2.5373,
        2.3588, 2.1755, 1.9886, 1.7991, 1.6075, 1.4145, 1.2204, 1.0255, 0.8299,
```

```
        0.6339, 0.4375, 0.2409])
----------------------------------------
iter  0  stage  3  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21,
21, 21, 21,
        21, 21, 21, 21, 21, 21,  0])
loss=  tensor(0.0413, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.9621,
5.9621, 5.9621, 5.9621, 5.4375, 5.0009, 4.6251, 4.2918, 3.9888,
        3.7076, 3.4421, 3.1881, 2.9426, 2.7034, 2.4688, 2.2377, 2.0092, 1.7826,
        1.5574, 1.3333, 1.1100, 0.8873, 0.6650, 0.4431, 0.2215]) return=
109925.7013290539
probs of actions:  tensor([0.9995, 0.9993, 0.9992, 0.9996, 0.9991, 0.9995,
0.9990, 0.9993, 0.9991,
        0.9991, 0.9994, 0.9994, 0.9993, 0.9993, 0.9993, 0.9994, 0.9990, 0.9990,
        0.9997, 1.0000, 0.9999, 1.0000, 1.0000, 0.9761, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4671, 0.5316, 0.5021, 0.4805, 0.4646, 0.4528, 0.4441,
0.4376, 0.4327,
        0.4291, 0.4264, 0.4244, 0.4228, 0.4217, 0.4208, 0.4202, 0.4197, 0.4194,
        0.4191, 0.4189, 0.4188, 0.4186, 0.4186, 0.4185, 0.4625])
finalReturns:  tensor([3.5297, 3.5738, 3.5457, 3.4688, 3.3579, 3.2234, 3.0719,
2.9084, 2.7359,
        2.5571, 2.3734, 2.1863, 1.9966, 1.8049, 1.6118, 1.4176, 1.2226, 1.0270,
        0.8309, 0.6345, 0.4379, 0.2411])
----------------------------------------
iter  0  stage  2  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21,
21, 21, 21,
        21, 21, 21, 21, 21, 21,  0])
loss=  tensor(0.0450, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.2565,
6.2565, 6.2565, 5.7103, 5.2590, 4.8729, 4.5324, 4.2242, 3.9392,
        3.6710, 3.4151, 3.1681, 2.9278, 2.6925, 2.4608, 2.2318, 2.0048, 1.7794,
        1.5551, 1.3317, 1.1089, 0.8865, 0.6646, 0.4429, 0.2214]) return=
109925.7013290539
probs of actions:  tensor([0.9995, 0.9993, 0.9992, 0.9996, 0.9991, 0.9995,
0.9990, 0.9993, 0.9991,
        0.9991, 0.9994, 0.9994, 0.9993, 0.9993, 0.9993, 0.9994, 0.9990, 0.9990,
        0.9997, 1.0000, 0.9999, 1.0000, 1.0000, 0.9760, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4671, 0.5316, 0.5021, 0.4805, 0.4646, 0.4528, 0.4441,
0.4376, 0.4327,
        0.4291, 0.4264, 0.4244, 0.4228, 0.4217, 0.4208, 0.4202, 0.4197, 0.4194,
        0.4191, 0.4189, 0.4188, 0.4186, 0.4186, 0.4185, 0.4625])
finalReturns:  tensor([3.7373, 3.7814, 3.7523, 3.6737, 3.5614, 3.4255, 3.2730,
3.1085, 2.9353,
        2.7559, 2.5718, 2.3844, 2.1944, 2.0025, 1.8093, 1.6150, 1.4199, 1.2242,
        1.0281, 0.8317, 0.6350, 0.4382, 0.2412])
----------------------------------------
```

```
iter  0  stage  1  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21,
21, 21, 21,
        21, 21, 21, 21, 21, 21,  0])
loss=  tensor(0.0484, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.5768,
6.5768, 6.0011, 5.5297, 5.1298, 4.7795, 4.4644, 4.1744, 3.9025,
        3.6439, 3.3950, 3.1533, 2.9168, 2.6843, 2.4547, 2.2273, 2.0016, 1.7770,
        1.5534, 1.3304, 1.1080, 0.8860, 0.6642, 0.4426, 0.2213]) return=
109925.7013290539
probs of actions:  tensor([0.9995, 0.9993, 0.9992, 0.9996, 0.9991, 0.9995,
0.9990, 0.9993, 0.9991,
        0.9991, 0.9994, 0.9994, 0.9993, 0.9993, 0.9993, 0.9994, 0.9990, 0.9990,
        0.9997, 1.0000, 0.9999, 1.0000, 1.0000, 0.9760, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4671, 0.5316, 0.5021, 0.4805, 0.4646, 0.4528, 0.4441,
0.4376, 0.4327,
        0.4291, 0.4264, 0.4244, 0.4228, 0.4217, 0.4208, 0.4202, 0.4197, 0.4194,
        0.4191, 0.4189, 0.4188, 0.4186, 0.4186, 0.4185, 0.4625])
finalReturns:  tensor([3.9487, 3.9928, 3.9620, 3.8815, 3.7671, 3.6294, 3.4754,
3.3097, 3.1356,
        2.9554, 2.7708, 2.5829, 2.3926, 2.2004, 2.0070, 1.8125, 1.6174, 1.4216,
        1.2254, 1.0290, 0.8323, 0.6354, 0.4384, 0.2413])
----------------------------------------
iter  0  stage  0  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21,
21, 21, 21,
        21, 21, 21, 21, 21, 21,  0])
loss=  tensor(0.0513, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.8278,
6.3166, 5.8178, 5.3989, 5.0355, 4.7109, 4.4142, 4.1374, 3.8753,
        3.6237, 3.3800, 3.1421, 2.9086, 2.6782, 2.4502, 2.2240, 1.9991, 1.7752,
        1.5521, 1.3295, 1.1074, 0.8855, 0.6639, 0.4425, 0.2212]) return=
109925.7013290539
probs of actions:  tensor([0.9995, 0.9993, 0.9992, 0.9996, 0.9991, 0.9995,
0.9990, 0.9993, 0.9991,
        0.9991, 0.9994, 0.9994, 0.9993, 0.9993, 0.9993, 0.9994, 0.9990, 0.9990,
        0.9997, 1.0000, 0.9999, 1.0000, 1.0000, 0.9759, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4671, 0.5316, 0.5021, 0.4805, 0.4646, 0.4528, 0.4441,
0.4376, 0.4327,
        0.4291, 0.4264, 0.4244, 0.4228, 0.4217, 0.4208, 0.4202, 0.4197, 0.4194,
        0.4191, 0.4189, 0.4188, 0.4186, 0.4186, 0.4185, 0.4625])
finalReturns:  tensor([4.1648, 4.2089, 4.1760, 4.0928, 3.9758, 3.8357, 3.6796,
3.5123, 3.3370,
        3.1558, 2.9704, 2.7819, 2.5911, 2.3987, 2.2050, 2.0103, 1.8150, 1.6191,
        1.4229, 1.2264, 1.0296, 0.8327, 0.6357, 0.4385, 0.2413])
0,[1e-05,1][1, 10000, 1, 1],1683106228 saved
[652587, 'tensor([0., 0., 0., 0., 0., 0., 0., 1., 0., 0., 0., 0.])',
109925.7013290539, 84773.34866887607, 0.051301419734954834, 1e-05, 1, 0,
```

```
'tensor([21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21,
21,\n        21, 21, 21, 21, 21, 21,  0])', '[1.   1.   1.   1.   1.   1.   1.
1.   1.   1.   1.   1.   1.   1.   1.\n 1.   1.   1.   1.   1.   1.   1.   1.
0.98 1.  ]', '0,[1e-05,1][1, 10000, 1, 1],1683106228', 25, 50,
167645.58347602683, 202386.15908071544, 82634.78314716663, 134766.63466666665,
131821.99466666667, 92498.06441395788, 92568.85885052304, 109925.70132905389,
109925.70132905389, 95781.68158887929, 92498.7981199441, 109927.0996773476]
policy reset
----------------------------------------
iter  1  stage  24  ep  99999   adversary:  AdversaryModes.fight_125
  actions:  tensor([0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,
0, 0, 0, 0,
        0])
loss=  tensor(-0., grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.4633,
0.4633, 0.4633, 0.4633, 0.4633, 0.4633, 0.4633, 0.4633, 0.4633,
        0.4633, 0.4633, 0.4633, 0.4633, 0.4633, 0.4633, 0.4633, 0.4633, 0.4633,
        0.4633, 0.4633, 0.4633, 0.4633, 0.4633, 0.4633, 0.4633]) return=
117778.10729910324
probs of actions:  tensor([0.9298, 0.9438, 0.0328, 0.9247, 0.9325, 0.9293,
0.9304, 0.9386, 0.9360,
        0.9318, 0.8983, 0.9275, 0.9239, 0.9333, 0.9266, 0.9286, 0.9126, 0.9336,
        0.0313, 0.8994, 0.9284, 0.9285, 0.9281, 0.9292, 0.9939],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5112, 0.4988, 0.4895, 0.4862, 0.4802, 0.4757, 0.4724,
0.4699, 0.4680,
        0.4666, 0.4655, 0.4648, 0.4642, 0.4637, 0.4634, 0.4631, 0.4630, 0.4628,
        0.4626, 0.4660, 0.4651, 0.4644, 0.4639, 0.4636, 0.4633])
finalReturns:  tensor([0.])
----------------------------------------
iter  1  stage  23  ep  99999   adversary:  AdversaryModes.fight_125
  actions:  tensor([17,  0,  0,  8,  1, 17,  9,  0,  6,  7,  9, 13,  3,  8, 10,
10,  8, 10,
        10,  8, 29,  8,  8, 10,  0])
loss=  tensor(0.0178, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.8227,
0.8227, 0.8227, 0.8227, 0.8227, 0.8227, 0.8227, 0.8227, 0.8227,
        0.8227, 0.8227, 0.8227, 0.8227, 0.8227, 0.8227, 0.8227, 0.8227, 0.8227,
        0.8227, 0.8227, 0.8227, 0.8227, 0.8227, 0.8227, 0.3931]) return=
110886.5992611681
probs of actions:  tensor([6.7410e-03, 3.9277e-01, 2.1554e-01, 1.2243e-01,
2.4300e-02, 4.6671e-03,
        1.0503e-01, 1.7088e-01, 4.1501e-02, 6.3581e-02, 9.4326e-02, 3.5382e-02,
        1.1196e-02, 1.0751e-01, 1.9669e-01, 2.2438e-01, 1.1883e-01, 2.5325e-01,
        2.4511e-01, 1.2646e-01, 7.9849e-05, 1.2561e-01, 1.2209e-01, 4.4844e-01,
        9.9323e-01], grad_fn=<ExpBackward0>)
rewards:  tensor([0.4823, 0.5606, 0.5352, 0.5101, 0.5313, 0.4884, 0.5573,
0.5210, 0.4544,
        0.4280, 0.4096, 0.3961, 0.4214, 0.3908, 0.3845, 0.3887, 0.3955, 0.3879,
        0.3913, 0.3975, 0.3153, 0.4585, 0.4382, 0.4196, 0.4251])
```

```
finalReturns:  tensor([0.0220, 0.0320])
----------------------------------------
iter  1  stage  22  ep  99999   adversary:  AdversaryModes.fight_125
  actions:  tensor([12, 15, 15, 15, 15, 12, 15, 15, 12, 13, 15, 15, 15, 15, 13,
15, 15, 15,
        15, 15, 15, 13, 15, 13,  0])
loss=  tensor(0.2627, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.3032,
1.3032, 1.3032, 1.3032, 1.3032, 1.3032, 1.3032, 1.3032, 1.3032,
        1.3032, 1.3032, 1.3032, 1.3032, 1.3032, 1.3032, 1.3032, 1.3032, 1.3032,
        1.3032, 1.3032, 1.3032, 1.3032, 1.3032, 0.8249, 0.3951]) return=
119446.69263091798
probs of actions:  tensor([0.0268, 0.6716, 0.6783, 0.6497, 0.7063, 0.0258,
0.6930, 0.7128, 0.0190,
        0.1823, 0.6454, 0.6303, 0.6423, 0.5961, 0.1839, 0.6566, 0.6457, 0.6843,
        0.7225, 0.6714, 0.6469, 0.2010, 0.8696, 0.1107, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.5196, 0.5546, 0.5312, 0.5140, 0.5094, 0.4811,
0.4768, 0.4817,
        0.4663, 0.4546, 0.4569, 0.4587, 0.4600, 0.4666, 0.4548, 0.4571, 0.4588,
        0.4601, 0.4611, 0.4618, 0.4680, 0.4558, 0.4635, 0.4750])
finalReturns:  tensor([0.0911, 0.1136, 0.0799])
----------------------------------------
iter  1  stage  21  ep  99999   adversary:  AdversaryModes.fight_125
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0023, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.6186,
1.6186, 1.6186, 1.6186, 1.6186, 1.6186, 1.6186, 1.6186, 1.6186,
        1.6186, 1.6186, 1.6186, 1.6186, 1.6186, 1.6186, 1.6186, 1.6186, 1.6186,
        1.6186, 1.6186, 1.6186, 1.6186, 1.1523, 0.7359, 0.3551]) return=
117051.59338141434
probs of actions:  tensor([0.9903, 0.9901, 0.9906, 0.9895, 0.9921, 0.9890,
0.9906, 0.9925, 0.9923,
        0.9895, 0.9863, 0.9882, 0.9871, 0.9857, 0.9883, 0.9888, 0.9872, 0.9912,
        0.9917, 0.9882, 0.9884, 0.9916, 0.9985, 0.9974, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5307, 0.5632, 0.5319, 0.5091, 0.4922, 0.4798,
0.4705, 0.4636,
        0.4585, 0.4547, 0.4518, 0.4497, 0.4481, 0.4469, 0.4460, 0.4453, 0.4448,
        0.4444, 0.4442, 0.4439, 0.4438, 0.4437, 0.4436, 0.4660])
finalReturns:  tensor([0.1785, 0.2010, 0.1737, 0.1109])
----------------------------------------
iter  1  stage  20  ep  40198   adversary:  AdversaryModes.fight_125
  actions:  tensor([15, 13, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0006, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.3391,
2.3391, 2.3391, 2.3391, 2.3391, 2.3391, 2.3391, 2.3391, 2.3391,
```

```
        2.3391, 2.3391, 2.3391, 2.3391, 2.3391, 2.3391, 2.3391, 2.3391, 2.3391,
        2.3391, 2.3391, 2.3391, 1.7892, 1.2936, 0.8369, 0.4082]) return=
132803.10474323918
probs of actions:  tensor([0.9989, 0.0011, 0.9989, 0.9988, 0.9992, 0.9988,
0.9990, 0.9993, 0.9992,
        0.9989, 0.9986, 0.9988, 0.9986, 0.9985, 0.9988, 0.9989, 0.9986, 0.9992,
        0.9991, 0.9987, 0.9991, 0.9995, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5363, 0.5556, 0.5413, 0.5436, 0.5394, 0.5364,
0.5341, 0.5323,
        0.5310, 0.5301, 0.5294, 0.5288, 0.5284, 0.5281, 0.5279, 0.5277, 0.5276,
        0.5275, 0.5274, 0.5273, 0.5273, 0.5273, 0.5273, 0.5497])
finalReturns:  tensor([0.3198, 0.3423, 0.3106, 0.2401, 0.1415])
----------------------------------------
iter  1  stage  19  ep  392   adversary:  AdversaryModes.fight_125
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0011, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.2790,
2.2790, 2.2790, 2.2790, 2.2790, 2.2790, 2.2790, 2.2790, 2.2790,
        2.2790, 2.2790, 2.2790, 2.2790, 2.2790, 2.2790, 2.2790, 2.2790, 2.2790,
        2.2790, 2.2790, 1.8123, 1.3957, 1.0148, 0.6595, 0.3229]) return=
117051.59338141434
probs of actions:  tensor([0.9990, 0.9990, 0.9990, 0.9989, 0.9992, 0.9988,
0.9990, 0.9992, 0.9992,
        0.9989, 0.9985, 0.9987, 0.9985, 0.9984, 0.9987, 0.9988, 0.9985, 0.9991,
        0.9991, 0.9990, 0.9990, 0.9994, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5307, 0.5632, 0.5319, 0.5091, 0.4922, 0.4798,
0.4705, 0.4636,
        0.4585, 0.4547, 0.4518, 0.4497, 0.4481, 0.4469, 0.4460, 0.4453, 0.4448,
        0.4444, 0.4442, 0.4439, 0.4438, 0.4437, 0.4436, 0.4660])
finalReturns:  tensor([0.4061, 0.4286, 0.4013, 0.3385, 0.2501, 0.1431])
----------------------------------------
iter  1  stage  18  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0019, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.5927,
2.5927, 2.5927, 2.5927, 2.5927, 2.5927, 2.5927, 2.5927, 2.5927,
        2.5927, 2.5927, 2.5927, 2.5927, 2.5927, 2.5927, 2.5927, 2.5927, 2.5927,
        2.5927, 2.1258, 1.7090, 1.3279, 0.9725, 0.6358, 0.3129]) return=
117051.59338141434
probs of actions:  tensor([0.9990, 0.9990, 0.9990, 0.9989, 0.9992, 0.9988,
0.9990, 0.9992, 0.9992,
        0.9989, 0.9985, 0.9987, 0.9985, 0.9984, 0.9987, 0.9988, 0.9985, 0.9991,
        0.9991, 0.9990, 0.9990, 0.9994, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
```

```
rewards:  tensor([0.4887, 0.5307, 0.5632, 0.5319, 0.5091, 0.4922, 0.4798,
0.4705, 0.4636,
        0.4585, 0.4547, 0.4518, 0.4497, 0.4481, 0.4469, 0.4460, 0.4453, 0.4448,
        0.4444, 0.4442, 0.4439, 0.4438, 0.4437, 0.4436, 0.4660])
finalReturns:  tensor([0.5368, 0.5593, 0.5320, 0.4692, 0.3808, 0.2738, 0.1531])
----------------------------------------
iter  1  stage  17  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0029, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.8994,
2.8994, 2.8994, 2.8994, 2.8994, 2.8994, 2.8994, 2.8994, 2.8994, 2.8994,
        2.8994, 2.8994, 2.8994, 2.8994, 2.8994, 2.8994, 2.8994, 2.8994, 2.8994,
        2.4321, 2.0150, 1.6337, 1.2782, 0.9414, 0.6184, 0.3054]) return=
117051.59338141434
probs of actions:  tensor([0.9990, 0.9990, 0.9990, 0.9989, 0.9992, 0.9988,
0.9990, 0.9992, 0.9992,
        0.9989, 0.9985, 0.9987, 0.9985, 0.9984, 0.9987, 0.9988, 0.9985, 0.9991,
        0.9991, 0.9990, 0.9990, 0.9994, 0.9999, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5307, 0.5632, 0.5319, 0.5091, 0.4922, 0.4798,
0.4705, 0.4636,
        0.4585, 0.4547, 0.4518, 0.4497, 0.4481, 0.4469, 0.4460, 0.4453, 0.4448,
        0.4444, 0.4442, 0.4439, 0.4438, 0.4437, 0.4436, 0.4660])
finalReturns:  tensor([0.6750, 0.6975, 0.6702, 0.6073, 0.5189, 0.4119, 0.2912,
0.1606])
----------------------------------------
iter  1  stage  16  ep  4983   adversary:  AdversaryModes.fight_125
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0028, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.2010,
3.2010, 3.2010, 3.2010, 3.2010, 3.2010, 3.2010, 3.2010, 3.2010, 3.2010,
        3.2010, 3.2010, 3.2010, 3.2010, 3.2010, 3.2010, 3.2010, 3.2010, 2.7332,
        2.3157, 1.9342, 1.5785, 1.2416, 0.9184, 0.6054, 0.2999]) return=
117051.59338141434
probs of actions:  tensor([0.9992, 0.9992, 0.9992, 0.9991, 0.9994, 0.9991,
0.9992, 0.9994, 0.9994,
        0.9991, 0.9988, 0.9990, 0.9988, 0.9988, 0.9990, 0.9991, 0.9990, 0.9998,
        0.9994, 0.9993, 0.9993, 0.9996, 1.0000, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5307, 0.5632, 0.5319, 0.5091, 0.4922, 0.4798,
0.4705, 0.4636,
        0.4585, 0.4547, 0.4518, 0.4497, 0.4481, 0.4469, 0.4460, 0.4453, 0.4448,
        0.4444, 0.4442, 0.4439, 0.4438, 0.4437, 0.4436, 0.4660])
finalReturns:  tensor([0.8187, 0.8412, 0.8139, 0.7510, 0.6625, 0.5555, 0.4348,
0.3041, 0.1661])
----------------------------------------
```

```
iter  1  stage  15  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0041, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.4991,
3.4991, 3.4991, 3.4991, 3.4991, 3.4991, 3.4991, 3.4991, 3.4991,
        3.4991, 3.4991, 3.4991, 3.4991, 3.4991, 3.4991, 3.4991, 3.0306, 2.6127,
        2.2308, 1.8748, 1.5378, 1.2145, 0.9014, 0.5958, 0.2958]) return=
117051.59338141434
probs of actions:  tensor([0.9992, 0.9992, 0.9992, 0.9991, 0.9994, 0.9991,
0.9992, 0.9994, 0.9994,
        0.9991, 0.9988, 0.9990, 0.9988, 0.9988, 0.9990, 0.9991, 0.9990, 0.9998,
        0.9994, 0.9993, 0.9993, 0.9996, 1.0000, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5307, 0.5632, 0.5319, 0.5091, 0.4922, 0.4798,
0.4705, 0.4636,
        0.4585, 0.4547, 0.4518, 0.4497, 0.4481, 0.4469, 0.4460, 0.4453, 0.4448,
        0.4444, 0.4442, 0.4439, 0.4438, 0.4437, 0.4436, 0.4660])
finalReturns:  tensor([0.9666, 0.9891, 0.9617, 0.8988, 0.8103, 0.7032, 0.5825,
0.4518, 0.3138,
        0.1702])
----------------------------------------
iter  1  stage  14  ep  20   adversary:  AdversaryModes.fight_125
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0055, grad_fn=<NegBackward0>)   ,  base rewards= tensor([3.7950,
3.7950, 3.7950, 3.7950, 3.7950, 3.7950, 3.7950, 3.7950, 3.7950,
        3.7950, 3.7950, 3.7950, 3.7950, 3.7950, 3.7950, 3.3256, 2.9070, 2.5247,
        2.1684, 1.8311, 1.5076, 1.1944, 0.8887, 0.5887, 0.2928]) return=
117051.59338141434
probs of actions:  tensor([0.9992, 0.9992, 0.9992, 0.9991, 0.9994, 0.9991,
0.9992, 0.9994, 0.9994,
        0.9991, 0.9988, 0.9990, 0.9988, 0.9988, 0.9990, 0.9991, 0.9990, 0.9998,
        0.9994, 0.9993, 0.9993, 0.9996, 1.0000, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5307, 0.5632, 0.5319, 0.5091, 0.4922, 0.4798,
0.4705, 0.4636,
        0.4585, 0.4547, 0.4518, 0.4497, 0.4481, 0.4469, 0.4460, 0.4453, 0.4448,
        0.4444, 0.4442, 0.4439, 0.4438, 0.4437, 0.4436, 0.4660])
finalReturns:  tensor([1.1176, 1.1401, 1.1127, 1.0497, 0.9612, 0.8541, 0.7333,
0.6026, 0.4645,
        0.3209, 0.1732])
----------------------------------------
iter  1  stage  13  ep  207   adversary:  AdversaryModes.fight_125
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
```

```
loss=  tensor(0.0071, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.0897,
4.0897, 4.0897, 4.0897, 4.0897, 4.0897, 4.0897, 4.0897, 4.0897,
        4.0897, 4.0897, 4.0897, 4.0897, 4.0897, 3.6191, 3.1997, 2.8167, 2.4600,
        2.1223, 1.7987, 1.4853, 1.1795, 0.8793, 0.5833, 0.2905]) return=
117051.59338141434
probs of actions:  tensor([0.9992, 0.9992, 0.9992, 0.9991, 0.9993, 0.9990,
0.9992, 0.9994, 0.9994,
        0.9991, 0.9988, 0.9990, 0.9988, 0.9990, 0.9990, 0.9992, 0.9990, 0.9998,
        0.9994, 0.9993, 0.9993, 0.9995, 1.0000, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5307, 0.5632, 0.5319, 0.5091, 0.4922, 0.4798,
0.4705, 0.4636,
        0.4585, 0.4547, 0.4518, 0.4497, 0.4481, 0.4469, 0.4460, 0.4453, 0.4448,
        0.4444, 0.4442, 0.4439, 0.4438, 0.4437, 0.4436, 0.4660])
finalReturns:  tensor([1.2710, 1.2935, 1.2661, 1.2030, 1.1144, 1.0072, 0.8865,
0.7557, 0.6176,
        0.4739, 0.3262, 0.1755])
---------------------------------------
iter  1  stage  12  ep  66   adversary:  AdversaryModes.fight_125
  actions:  tensor([15, 13, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 13, 15, 15, 15,  0])
loss=  tensor(6.1737, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.2969,
5.2969, 5.2969, 5.2969, 5.2969, 5.2969, 5.2969, 5.2969, 5.2969,
        5.2969, 5.2969, 5.2969, 5.2969, 4.7456, 4.2490, 3.7915, 3.3623, 2.9537,
        2.5602, 2.1779, 1.8039, 1.4360, 1.0727, 0.7128, 0.3554]) return=
132656.87375173956
probs of actions:  tensor([9.9921e-01, 8.1956e-04, 9.9919e-01, 9.9913e-01,
9.9937e-01, 9.9911e-01,
        9.9927e-01, 9.9944e-01, 9.9943e-01, 9.9919e-01, 9.9892e-01, 9.9912e-01,
        9.9905e-01, 9.9916e-01, 9.9913e-01, 9.9935e-01, 9.9912e-01, 9.9982e-01,
        9.9947e-01, 9.9936e-01, 6.5138e-04, 9.9961e-01, 9.9997e-01, 9.9984e-01,
        1.0000e+00], grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5363, 0.5556, 0.5413, 0.5436, 0.5394, 0.5364,
0.5341, 0.5323,
        0.5310, 0.5301, 0.5294, 0.5288, 0.5284, 0.5281, 0.5279, 0.5277, 0.5276,
        0.5275, 0.5274, 0.5329, 0.5199, 0.5217, 0.5231, 0.5466])
finalReturns:  tensor([1.5707, 1.5932, 1.5614, 1.4908, 1.3922, 1.2731, 1.1390,
0.9938, 0.8404,
        0.6753, 0.5187, 0.3569, 0.1912])
---------------------------------------
iter  1  stage  11  ep  19   adversary:  AdversaryModes.fight_125
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0115, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.6792,
4.6792, 4.6792, 4.6792, 4.6792, 4.6792, 4.6792, 4.6792, 4.6792,
        4.6792, 4.6792, 4.6792, 4.2048, 3.7828, 3.3979, 3.0398, 2.7012, 2.3768,
```

```
        2.0628, 1.7566, 1.4562, 1.1600, 0.8670, 0.5764, 0.2875]) return=
117051.59338141434
probs of actions:  tensor([0.9992, 0.9992, 0.9992, 0.9991, 0.9993, 0.9990,
0.9992, 0.9994, 0.9993,
        0.9991, 0.9987, 0.9990, 0.9989, 0.9990, 0.9990, 0.9992, 0.9990, 0.9998,
        0.9994, 0.9992, 0.9992, 0.9995, 1.0000, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5307, 0.5632, 0.5319, 0.5091, 0.4922, 0.4798,
0.4705, 0.4636,
        0.4585, 0.4547, 0.4518, 0.4497, 0.4481, 0.4469, 0.4460, 0.4453, 0.4448,
        0.4444, 0.4442, 0.4439, 0.4438, 0.4437, 0.4436, 0.4660])
finalReturns:  tensor([1.5830, 1.6055, 1.5779, 1.5147, 1.4259, 1.3185, 1.1976,
1.0667, 0.9285,
        0.7848, 0.6370, 0.4862, 0.3332, 0.1785])
----------------------------------------
iter  1  stage  10  ep  178   adversary:  AdversaryModes.fight_125
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0126, grad_fn=<NegBackward0>)   ,  base rewards= tensor([4.9759,
4.9759, 4.9759, 4.9759, 4.9759, 4.9759, 4.9759, 4.9759, 4.9759, 4.9759,
        4.9759, 4.9759, 4.4987, 4.0746, 3.6883, 3.3292, 2.9898, 2.6648, 2.3505,
        2.0440, 1.7433, 1.4469, 1.1538, 0.8631, 0.5741, 0.2866]) return=
117051.59338141434
probs of actions:  tensor([0.9992, 0.9992, 0.9992, 0.9991, 0.9994, 0.9991,
0.9992, 0.9994, 0.9994,
        0.9991, 0.9990, 0.9992, 0.9991, 0.9992, 0.9990, 0.9993, 0.9990, 0.9998,
        0.9994, 0.9993, 0.9992, 0.9996, 1.0000, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5307, 0.5632, 0.5319, 0.5091, 0.4922, 0.4798,
0.4705, 0.4636,
        0.4585, 0.4547, 0.4518, 0.4497, 0.4481, 0.4469, 0.4460, 0.4453, 0.4448,
        0.4444, 0.4442, 0.4439, 0.4438, 0.4437, 0.4436, 0.4660])
finalReturns:  tensor([1.7410, 1.7635, 1.7358, 1.6724, 1.5834, 1.4759, 1.3549,
1.2239, 1.0856,
        0.9419, 0.7941, 0.6432, 0.4902, 0.3354, 0.1794])
----------------------------------------
iter  1  stage  9  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0149, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.2754,
5.2754, 5.2754, 5.2754, 5.2754, 5.2754, 5.2754, 5.2754, 5.2754, 5.2754,
        5.2754, 4.7944, 4.3676, 3.9793, 3.6188, 3.2784, 2.9527, 2.6378, 2.3308,
        2.0299, 1.7333, 1.4400, 1.1491, 0.8601, 0.5725, 0.2859]) return=
117051.59338141434
probs of actions:  tensor([0.9992, 0.9992, 0.9992, 0.9991, 0.9994, 0.9991,
0.9992, 0.9994, 0.9994,
```

```
        0.9991, 0.9990, 0.9992, 0.9991, 0.9992, 0.9990, 0.9993, 0.9990, 0.9998,
        0.9994, 0.9993, 0.9992, 0.9996, 1.0000, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5307, 0.5632, 0.5319, 0.5091, 0.4922, 0.4798,
0.4705, 0.4636,
        0.4585, 0.4547, 0.4518, 0.4497, 0.4481, 0.4469, 0.4460, 0.4453, 0.4448,
        0.4444, 0.4442, 0.4439, 0.4438, 0.4437, 0.4436, 0.4660])
finalReturns:  tensor([1.9000, 1.9225, 1.8946, 1.8310, 1.7419, 1.6342, 1.5130,
1.3820, 1.2435,
        1.0997, 0.9519, 0.8010, 0.6479, 0.4931, 0.3371, 0.1802])
---------------------------------------
iter  1  stage  8  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0167, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.5790,
5.5790, 5.5790, 5.5790, 5.5790, 5.5790, 5.5790, 5.5790, 5.5790,
        5.0929, 4.6624, 4.2716, 3.9092, 3.5674, 3.2407, 2.9251, 2.6176, 2.3162,
        2.0193, 1.7258, 1.4348, 1.1457, 0.8579, 0.5712, 0.2853]) return=
117051.59338141434
probs of actions:  tensor([0.9992, 0.9992, 0.9992, 0.9991, 0.9994, 0.9991,
0.9992, 0.9994, 0.9994,
        0.9991, 0.9990, 0.9992, 0.9991, 0.9992, 0.9990, 0.9993, 0.9990, 0.9998,
        0.9994, 0.9993, 0.9992, 0.9996, 1.0000, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5307, 0.5632, 0.5319, 0.5091, 0.4922, 0.4798,
0.4705, 0.4636,
        0.4585, 0.4547, 0.4518, 0.4497, 0.4481, 0.4469, 0.4460, 0.4453, 0.4448,
        0.4444, 0.4442, 0.4439, 0.4438, 0.4437, 0.4436, 0.4660])
finalReturns:  tensor([2.0601, 2.0826, 2.0545, 1.9906, 1.9012, 1.7933, 1.6719,
1.5406, 1.4021,
        1.2582, 1.1102, 0.9593, 0.8062, 0.6514, 0.4953, 0.3384, 0.1807])
---------------------------------------
iter  1  stage  7  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 13, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(9.7105, grad_fn=<NegBackward0>)   ,  base rewards= tensor([5.8884,
5.8884, 5.8884, 5.8884, 5.8884, 5.8884, 5.8884, 5.8884, 5.3954,
        4.9602, 4.5659, 4.2010, 3.8574, 3.5293, 3.2127, 2.9045, 2.6026, 2.3053,
        2.0115, 1.7202, 1.4309, 1.1430, 0.8563, 0.5703, 0.2849]) return=
116862.23949259547
probs of actions:  tensor([9.9923e-01, 9.9920e-01, 9.9920e-01, 9.9914e-01,
9.9936e-01, 9.9906e-01,
        9.9921e-01, 9.9939e-01, 9.9938e-01, 9.9910e-01, 9.9900e-01, 9.9920e-01,
        9.9909e-01, 9.9916e-01, 9.9903e-01, 9.9927e-01, 9.7531e-04, 9.9980e-01,
        9.9940e-01, 9.9928e-01, 9.9924e-01, 9.9955e-01, 9.9997e-01, 9.9983e-01,
        1.0000e+00], grad_fn=<ExpBackward0>)
```

```
rewards:  tensor([0.4887, 0.5307, 0.5632, 0.5319, 0.5091, 0.4922, 0.4798,
0.4705, 0.4636,
        0.4585, 0.4547, 0.4518, 0.4497, 0.4481, 0.4469, 0.4460, 0.4509, 0.4380,
        0.4393, 0.4403, 0.4411, 0.4416, 0.4420, 0.4424, 0.4651])
finalReturns:  tensor([2.2022, 2.2247, 2.1963, 2.1321, 2.0423, 1.9341, 1.8124,
1.6810, 1.5423,
        1.3982, 1.2446, 1.1004, 0.9523, 0.8013, 0.6481, 0.4932, 0.3372, 0.1802])
----------------------------------------
iter  1  stage  6  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0211, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.2060,
6.2060, 6.2060, 6.2060, 6.2060, 6.2060, 6.2060, 5.7037, 5.2620,
        4.8630, 4.4948, 4.1487, 3.8189, 3.5010, 3.1918, 2.8891, 2.5913, 2.2971,
        2.0056, 1.7160, 1.4280, 1.1411, 0.8550, 0.5696, 0.2846]) return=
117051.59338141434
probs of actions:  tensor([0.9992, 0.9992, 0.9992, 0.9991, 0.9994, 0.9991,
0.9992, 0.9994, 0.9994,
        0.9991, 0.9990, 0.9992, 0.9991, 0.9992, 0.9990, 0.9993, 0.9990, 0.9998,
        0.9994, 0.9993, 0.9992, 0.9996, 1.0000, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5307, 0.5632, 0.5319, 0.5091, 0.4922, 0.4798,
0.4705, 0.4636,
        0.4585, 0.4547, 0.4518, 0.4497, 0.4481, 0.4469, 0.4460, 0.4453, 0.4448,
        0.4444, 0.4442, 0.4439, 0.4438, 0.4437, 0.4436, 0.4660])
finalReturns:  tensor([2.3834, 2.4059, 2.3771, 2.3124, 2.2221, 2.1135, 1.9915,
1.8597, 1.7208,
        1.5766, 1.4284, 1.2773, 1.1240, 0.9691, 0.8130, 0.6559, 0.4982, 0.3400,
        0.1814])
----------------------------------------
iter  1  stage  5  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0240, grad_fn=<NegBackward0>)   ,  base rewards= tensor([6.5346,
6.5346, 6.5346, 6.5346, 6.5346, 6.5346, 6.0199, 5.5694, 5.1642,
        4.7915, 4.4422, 4.1100, 3.7903, 3.4798, 3.1762, 2.8777, 2.5829, 2.2910,
        2.0012, 1.7129, 1.4258, 1.1396, 0.8541, 0.5691, 0.2844]) return=
117051.59338141434
probs of actions:  tensor([0.9992, 0.9992, 0.9992, 0.9991, 0.9994, 0.9991,
0.9992, 0.9994, 0.9994,
        0.9991, 0.9990, 0.9992, 0.9991, 0.9992, 0.9990, 0.9993, 0.9990, 0.9998,
        0.9994, 0.9993, 0.9992, 0.9996, 1.0000, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5307, 0.5632, 0.5319, 0.5091, 0.4922, 0.4798,
0.4705, 0.4636,
        0.4585, 0.4547, 0.4518, 0.4497, 0.4481, 0.4469, 0.4460, 0.4453, 0.4448,
```

```
        0.4444, 0.4442, 0.4439, 0.4438, 0.4437, 0.4436, 0.4660])
finalReturns:  tensor([2.5469, 2.5694, 2.5402, 2.4748, 2.3839, 2.2747, 2.1522,
2.0201, 1.8809,
        1.7364, 1.5880, 1.4368, 1.2834, 1.1284, 0.9722, 0.8152, 0.6574, 0.4992,
        0.3405, 0.1816])
---------------------------------------
iter  1  stage  4  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([15, 13, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0261, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.1639,
8.1639, 8.1639, 8.1639, 8.1639, 7.5979, 7.0908, 6.6257, 6.1910,
        5.7784, 5.3820, 4.9976, 4.6219, 4.2528, 3.8886, 3.5281, 3.1702, 2.8144,
        2.4601, 2.1069, 1.7546, 1.4030, 1.0518, 0.7009, 0.3504]) return=
132803.10474323918
probs of actions:  tensor([9.9923e-01, 8.0644e-04, 9.9920e-01, 9.9914e-01,
9.9938e-01, 9.9912e-01,
        9.9928e-01, 9.9945e-01, 9.9944e-01, 9.9920e-01, 9.9912e-01, 9.9930e-01,
        9.9920e-01, 9.9926e-01, 9.9915e-01, 9.9936e-01, 9.9914e-01, 9.9982e-01,
        9.9948e-01, 9.9937e-01, 9.9934e-01, 9.9961e-01, 9.9997e-01, 9.9985e-01,
        1.0000e+00], grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5363, 0.5556, 0.5413, 0.5436, 0.5394, 0.5364,
0.5341, 0.5323,
        0.5310, 0.5301, 0.5294, 0.5288, 0.5284, 0.5281, 0.5279, 0.5277, 0.5276,
        0.5275, 0.5274, 0.5273, 0.5273, 0.5273, 0.5273, 0.5497])
finalReturns:  tensor([2.9945, 3.0170, 2.9847, 2.9134, 2.8140, 2.6943, 2.5596,
2.4140, 2.2603,
        2.1006, 1.9364, 1.7689, 1.5988, 1.4270, 1.2537, 1.0794, 0.9043, 0.7286,
        0.5525, 0.3761, 0.1994])
---------------------------------------
iter  1  stage  3  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0295, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.2430,
7.2430, 7.2430, 7.2430, 6.6886, 6.2103, 5.7855, 5.3986, 5.0391,
        4.6994, 4.3742, 4.0597, 3.7530, 3.4523, 3.1559, 2.8627, 2.5719, 2.2830,
        1.9954, 1.7088, 1.4230, 1.1377, 0.8529, 0.5684, 0.2841]) return=
117051.59338141434
probs of actions:  tensor([0.9992, 0.9992, 0.9992, 0.9991, 0.9993, 0.9991,
0.9992, 0.9994, 0.9994,
        0.9991, 0.9990, 0.9992, 0.9991, 0.9992, 0.9990, 0.9993, 0.9990, 0.9998,
        0.9994, 0.9993, 0.9992, 0.9995, 1.0000, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5307, 0.5632, 0.5319, 0.5091, 0.4922, 0.4798,
0.4705, 0.4636,
        0.4585, 0.4547, 0.4518, 0.4497, 0.4481, 0.4469, 0.4460, 0.4453, 0.4448,
        0.4444, 0.4442, 0.4439, 0.4438, 0.4437, 0.4436, 0.4660])
```

```
finalReturns:  tensor([2.8795, 2.9020, 2.8712, 2.8039, 2.7110, 2.6000, 2.4761,
2.3427, 2.2026,
        2.0573, 1.9084, 1.7567, 1.6030, 1.4478, 1.2914, 1.1342, 0.9763, 0.8180,
        0.6593, 0.5004, 0.3412, 0.1819])
----------------------------------------
iter  1  stage  2  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0325, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.6361,
7.6361, 7.6361, 7.0504, 6.5504, 6.1103, 5.7124, 5.3450, 4.9996,
        4.6702, 4.3525, 4.0436, 3.7411, 3.4435, 3.1493, 2.8579, 2.5684, 2.2804,
        1.9935, 1.7075, 1.4220, 1.1371, 0.8525, 0.5681, 0.2840]) return=
117051.59338141434
probs of actions:  tensor([0.9992, 0.9992, 0.9992, 0.9991, 0.9993, 0.9991,
0.9992, 0.9994, 0.9994,
        0.9991, 0.9990, 0.9992, 0.9991, 0.9992, 0.9990, 0.9993, 0.9990, 0.9998,
        0.9994, 0.9993, 0.9992, 0.9995, 1.0000, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5307, 0.5632, 0.5319, 0.5091, 0.4922, 0.4798,
0.4705, 0.4636,
        0.4585, 0.4547, 0.4518, 0.4497, 0.4481, 0.4469, 0.4460, 0.4453, 0.4448,
        0.4444, 0.4442, 0.4439, 0.4438, 0.4437, 0.4436, 0.4660])
finalReturns:  tensor([3.0496, 3.0721, 3.0402, 2.9713, 2.8769, 2.7646, 2.6395,
2.5053, 2.3644,
        2.2186, 2.0692, 1.9172, 1.7633, 1.6078, 1.4513, 1.2940, 1.1361, 0.9777,
        0.8189, 0.6599, 0.5008, 0.3414, 0.1820])
----------------------------------------
iter  1  stage  1  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 13, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(9.9530, grad_fn=<NegBackward0>)   ,  base rewards= tensor([7.9931,
7.9931, 7.4399, 6.9102, 6.4492, 6.0365, 5.6584, 5.3053, 4.9702,
        4.6484, 4.3364, 4.0316, 3.7322, 3.4369, 3.1444, 2.8543, 2.5657, 2.2785,
        1.9921, 1.7065, 1.4214, 1.1366, 0.8522, 0.5680, 0.2839]) return=
116862.23949259547
probs of actions:  tensor([9.9922e-01, 9.9919e-01, 9.9919e-01, 9.9913e-01,
9.9935e-01, 9.9905e-01,
        9.9921e-01, 9.9939e-01, 9.9937e-01, 9.9910e-01, 9.9899e-01, 9.9919e-01,
        9.9908e-01, 9.9915e-01, 9.9902e-01, 9.9926e-01, 9.9577e-04, 9.9979e-01,
        9.9939e-01, 9.9927e-01, 9.9923e-01, 9.9955e-01, 9.9997e-01, 9.9982e-01,
        1.0000e+00], grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5307, 0.5632, 0.5319, 0.5091, 0.4922, 0.4798,
0.4705, 0.4636,
        0.4585, 0.4547, 0.4518, 0.4497, 0.4481, 0.4469, 0.4460, 0.4509, 0.4380,
        0.4393, 0.4403, 0.4411, 0.4416, 0.4420, 0.4424, 0.4651])
finalReturns:  tensor([3.2044, 3.2269, 3.1934, 3.1225, 3.0261, 2.9120, 2.7854,
```

```
2.6499, 2.5081,
        2.3616, 2.2116, 2.0592, 1.9049, 1.7492, 1.5925, 1.4350, 1.2714, 1.1197,
        0.9660, 0.8108, 0.6545, 0.4973, 0.3395, 0.1812])
---------------------------------------
iter  1  stage  0  ep  0   adversary:  AdversaryModes.fight_125
  actions:  tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15, 15, 15,
        15, 15, 15, 15, 15, 15,  0])
loss=  tensor(0.0391, grad_fn=<NegBackward0>)   ,  base rewards= tensor([8.3070,
7.7957, 7.2969, 6.8074, 6.3746, 5.9820, 5.6184, 5.2757, 4.9483,
        4.6322, 4.3243, 4.0227, 3.7256, 3.4319, 3.1408, 2.8516, 2.5638, 2.2770,
        1.9911, 1.7057, 1.4208, 1.1363, 0.8520, 0.5679, 0.2839]) return=
117051.59338141434
probs of actions:  tensor([0.9992, 0.9992, 0.9992, 0.9991, 0.9993, 0.9990,
0.9992, 0.9994, 0.9994,
        0.9991, 0.9990, 0.9992, 0.9991, 0.9991, 0.9990, 0.9993, 0.9990, 0.9998,
        0.9994, 0.9993, 0.9992, 0.9995, 1.0000, 0.9998, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.5307, 0.5632, 0.5319, 0.5091, 0.4922, 0.4798,
0.4705, 0.4636,
        0.4585, 0.4547, 0.4518, 0.4497, 0.4481, 0.4469, 0.4460, 0.4453, 0.4448,
        0.4444, 0.4442, 0.4439, 0.4438, 0.4437, 0.4436, 0.4660])
finalReturns:  tensor([3.3982, 3.4207, 3.3888, 3.3152, 3.2161, 3.0996, 2.9710,
2.8339, 2.6908,
        2.5433, 2.3926, 2.2396, 2.0848, 1.9288, 1.7718, 1.6141, 1.4559, 1.2974,
        1.1385, 0.9794, 0.8201, 0.6607, 0.5013, 0.3417, 0.1821])
0,[1e-05,1][1, 10000, 1, 1],1683118086 saved
[526084, 'tensor([0., 0., 0., 0., 0., 0., 0., 1., 0., 0., 0., 0.])',
117051.59338141434, 92893.76204687786, 0.03911364823579788, 1e-05, 1, 0,
'tensor([15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,
15,\n        15, 15, 15, 15, 15, 15,  0])', '[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.\n 1.]', '0,[1e-05,1][1, 10000, 1,
1],1683118086', 25, 50, 161287.58344151574, 184572.3091501231,
73445.38232037451, 135545.41866666666, 132647.75466666667, 112524.66365021843,
112524.66365021843, 117051.59338141435, 117034.7937498715, 85375.85726043607,
112524.66365021843, 117051.59338141435]
policy reset
---------------------------------------
iter  2  stage  24  ep  99999   adversary:  AdversaryModes.fight_125
  actions:  tensor([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 3, 0, 0,
0, 0, 0, 0,
        0])
loss=  tensor(-0., grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.4645,
0.4645, 0.4645, 0.4645, 0.4645, 0.4645, 0.4645, 0.4645, 0.4645,
        0.4645, 0.4645, 0.4645, 0.4645, 0.4645, 0.4645, 0.4645, 0.4645, 0.4645,
        0.4645, 0.4645, 0.4645, 0.4645, 0.4645, 0.4645, 0.4645]) return=
118139.09703589762
probs of actions:  tensor([0.9190, 0.8833, 0.9029, 0.8854, 0.9188, 0.9049,
```

```
0.8844, 0.9179, 0.8879,
        0.9124, 0.8865, 0.0176, 0.8999, 0.9145, 0.9051, 0.8961, 0.8862, 0.0062,
        0.8904, 0.8916, 0.9037, 0.8957, 0.9033, 0.8869, 0.9912],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5112, 0.4988, 0.4896, 0.4827, 0.4776, 0.4738, 0.4709,
0.4688, 0.4672,
        0.4660, 0.4651, 0.4640, 0.4707, 0.4687, 0.4671, 0.4659, 0.4650, 0.4635,
        0.4742, 0.4712, 0.4690, 0.4673, 0.4661, 0.4652, 0.4645])
finalReturns:  tensor([0.])
```

```
---------------------------------------------------------------------------
KeyboardInterrupt                         Traceback (most recent call last)
Input In [3], in <cell line: 1>()
      7 neuralNet=NNBase(num_input=game.T+2+game.advHistoryNum,␣
 ↪lr=hyperParams[0],num_actions=50)
      8 algorithm = ReinforceAlgorithm(game, neuralNet, numberIterations=3,␣
 ↪numberEpisodes=3_000_000, discountFactor =hyperParams[1])
---> 11␣
 ↪algorithm.solver(print_step=100_000,options=codeParams,converge_break=True)

File ~\Documents\EquiLearn\PGM_base\learningBase.py:136, in ReinforceAlgorithm.
 ↪solver(self, print_step, options, converge_break)
    133 self.loss.append([])
    135 for stage in range(self.env.T-1, -1, -1):
--> 136     ␣
 ↪self.learn_stage_onwards(iter,stage=stage, episodes=int(self.numberEpisodes/s lf.env.T), pr
    137␣
 ↪                    prob_break_limit_ln=(self.probBreakLn if converge_break e se None), wri
    139 axs[iter][0].scatter(range(len(self.returns[iter])), self.returns[iter]
    140 axs[iter][1].scatter(range(len(self.loss[iter])), self.loss[iter])

File ~\Documents\EquiLearn\PGM_base\learningBase.py:200, in ReinforceAlgorithm.
 ↪learn_stage_onwards(self, iter, stage, episodes, print_step,␣
 ↪prob_break_limit_ln, options, lr, just_stage, write_save)
    198 actionsLogProbs = action_logprobs[stage:]
    199 discRewards = self.returnsComputation(rewards=rewards)
--> 200 baseRewards = self.computeBase(
    201␣
 ↪    self.env.prices[1], initDemand=self.env.demandPotential[0][stage], startS age=stage)/
 ↪options[1]
    202 baseDiscReturns = discRewards-baseRewards
    203 finalReturns = baseDiscReturns[stage:]

File ~\Documents\EquiLearn\PGM_base\learningBase.py:299, in ReinforceAlgorithm.
 ↪computeBase(self, advPrices, startStage, initDemand)
    297     profit[i] = (demand-price)*(price-self.env.costs[0])
    298     demand += (advPrices[i]-price)/2
--> 299 return self.returnsComputation(rewards=profit)
```

```
File ~\Documents\EquiLearn\PGM_base\learningBase.py:84, in ReinforceAlgorithm.
 ↪returnsComputation(self, rewards, episodeMemory)
     82 discRewards[-1] = rewards[-1]
     83 for i in range(len(rewards)-2, -1, -1):
---> 84     discRewards[i] = rewards[i] + self.gamma*discRewards[i+1]
     85 return discRewards

KeyboardInterrupt:
```

[ ]: