small

May 4, 2023

[1]:
```
from learningBase import ReinforceAlgorithm
import environmentModelBase as model
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]:
```
const95= model.Strategy(model.StrategyType.static, model.const,name="const95",␣
 ↪firstPrice=95 )
actionStep=3
adv= model.Adversary()
adv._strategies.append(const95)
adv._strategyProbs=torch.ones(1)
game = model.Model(totalDemand = 400,
               tupleCosts = (57, 71),
              totalStages = 25, adversary=adv, stateAdvHistory=1,␣
 ↪actionStep=actionStep)
```

[3]:
```
hyperParams=[0.000005, 1, 0]
codeParams=[1, 10000, 1, 1]
```

[4]:
```
neuralNet=NNBase(num_input=game.T+2+game.stateAdvHistory,␣
 ↪lr=hyperParams[0],num_actions=18,action_step=actionStep)
algorithm = ReinforceAlgorithm(game, neuralNet, numberIterations=2,␣
 ↪numberEpisodes=50_000, discountFactor =hyperParams[1])


algorithm.solver(print_step=30_000,options=codeParams,converge_break=True)
```

```
policy reset
----------------------------------------
iter  0  stage  24  ep  29999   adversary:  const95-1.0,
  actions:  tensor([ 3, 42,  3, 12,  0,  0,  0,  6,  6, 36,  6, 15,  6, 30,  9,
0,  3,  6,
         3, 42,  3,  6,  6,  3,  0])
loss=  tensor(-0., grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.2113,
```

```
0.2113, 0.2113, 0.2113, 0.2113, 0.2113, 0.2113, 0.2113, 0.2113,
        0.2113, 0.2113, 0.2113, 0.2113, 0.2113, 0.2113, 0.2113, 0.2113, 0.2113,
        0.2113, 0.2113, 0.2113, 0.2113, 0.2113, 0.2113, 0.2113]) return=
62979.42909757305
probs of actions:  tensor([0.2569, 0.0078, 0.2615, 0.0434, 0.2830, 0.2847,
0.2901, 0.1838, 0.1792,
        0.0102, 0.1705, 0.0234, 0.1853, 0.0105, 0.0954, 0.2553, 0.2298, 0.2041,
        0.2421, 0.0096, 0.2458, 0.1886, 0.1877, 0.2351, 0.3951],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5103, 0.2316, 0.4602, 0.3599, 0.3409, 0.2840, 0.2447,
0.2136, 0.2075,
        0.0771, 0.2730, 0.2320, 0.2574, 0.1532, 0.2833, 0.2729, 0.2361, 0.2150,
        0.2113, 0.0243, 0.2864, 0.2509, 0.2349, 0.2260, 0.2113])
finalReturns:  tensor([0.])
----------------------------------------
iter  0  stage  23  ep  29999   adversary:  const95-1.0,
  actions:  tensor([3, 0, 0, 3, 0, 0, 9, 0, 6, 3, 0, 0, 0, 6, 0, 0, 3, 6, 3, 0,
3, 6, 3, 3,
        3])
loss=  tensor(0.0080, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.3337,
0.3337, 0.3337, 0.3337, 0.3337, 0.3337, 0.3337, 0.3337, 0.3337,
        0.3337, 0.3337, 0.3337, 0.3337, 0.3337, 0.3337, 0.3337, 0.3337, 0.3337,
        0.3337, 0.3337, 0.3337, 0.3337, 0.3337, 0.3337, 0.1635]) return=
52462.04122834946
probs of actions:  tensor([0.4363, 0.2692, 0.2836, 0.4190, 0.2666, 0.2884,
0.0413, 0.2971, 0.2062,
        0.4231, 0.2851, 0.2973, 0.2680, 0.2101, 0.2771, 0.2523, 0.4025, 0.2236,
        0.4208, 0.2826, 0.4316, 0.2136, 0.4160, 0.4585, 0.4117],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5103, 0.4080, 0.3295, 0.2753, 0.2467, 0.2186, 0.1905,
0.2040, 0.1846,
        0.1886, 0.1840, 0.1737, 0.1661, 0.1569, 0.1685, 0.1623, 0.1568, 0.1567,
        0.1674, 0.1683, 0.1612, 0.1600, 0.1700, 0.1693, 0.1688])
finalReturns:  tensor([0.0043, 0.0052])
----------------------------------------
iter  0  stage  23  ep  59999   adversary:  const95-1.0,
  actions:  tensor([ 6,  6,  0,  6,  6,  3,  0,  3, 27,  0,  6,  3,  0,  3,  3,
6,  3,  3,
         0,  6,  3,  3,  6,  3,  0])
loss=  tensor(0.0133, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.3451,
0.3451, 0.3451, 0.3451, 0.3451, 0.3451, 0.3451, 0.3451, 0.3451,
        0.3451, 0.3451, 0.3451, 0.3451, 0.3451, 0.3451, 0.3451, 0.3451, 0.3451,
        0.3451, 0.3451, 0.3451, 0.3451, 0.3451, 0.3451, 0.1684]) return=
55844.69544871044
probs of actions:  tensor([0.3940, 0.4652, 0.2322, 0.4760, 0.4480, 0.2834,
0.1864, 0.2675, 0.0010,
        0.1776, 0.4118, 0.2668, 0.1869, 0.2454, 0.2470, 0.4796, 0.2581, 0.2561,
        0.2266, 0.4460, 0.2829, 0.2552, 0.4303, 0.1324, 0.6613],
```

```
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5076, 0.4140, 0.3537, 0.2891, 0.2624, 0.2459, 0.2257,
0.2028, 0.1216,
        0.2433, 0.2126, 0.2095, 0.1994, 0.1839, 0.1797, 0.1738, 0.1805, 0.1771,
        0.1755, 0.1638, 0.1729, 0.1715, 0.1677, 0.1758, 0.1746])
finalReturns:  tensor([0.0053, 0.0062])
----------------------------------------
iter  0  stage  23  ep  89999   adversary:  const95-1.0,
  actions:  tensor([6, 6, 9, 6, 6, 6, 6, 3, 6, 6, 9, 9, 6, 6, 6, 6, 9, 6, 6, 6,
6, 6, 0, 9,
        0])
loss=  tensor(0.0300, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.3550,
0.3550, 0.3550, 0.3550, 0.3550, 0.3550, 0.3550, 0.3550, 0.3550,
        0.3550, 0.3550, 0.3550, 0.3550, 0.3550, 0.3550, 0.3550, 0.3550, 0.3550,
        0.3550, 0.3550, 0.3550, 0.3550, 0.3550, 0.3550, 0.1725]) return=
59413.20879318385
probs of actions:  tensor([0.4593, 0.5525, 0.0732, 0.5638, 0.5460, 0.4904,
0.5642, 0.1712, 0.5676,
        0.5611, 0.0841, 0.0820, 0.5512, 0.5398, 0.5214, 0.5633, 0.0787, 0.5516,
        0.4780, 0.5224, 0.4695, 0.5310, 0.2116, 0.0809, 0.8957],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5076, 0.4140, 0.3456, 0.3139, 0.2801, 0.2560, 0.2386,
0.2286, 0.2097,
        0.2047, 0.1964, 0.2005, 0.2080, 0.2034, 0.2000, 0.1975, 0.1911, 0.2009,
        0.1982, 0.1961, 0.1946, 0.1934, 0.1962, 0.1744, 0.1917])
finalReturns:  tensor([0.0111, 0.0192])
----------------------------------------
iter  0  stage  22  ep  29999   adversary:  const95-1.0,
  actions:  tensor([15,  9,  6,  6,  0,  6,  9,  6,  6,  9,  9,  9,  6,  6,  9,
6,  6,  9,
        12,  6,  9,  6,  9,  6,  0])
loss=  tensor(0.0575, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.5794,
0.5794, 0.5794, 0.5794, 0.5794, 0.5794, 0.5794, 0.5794, 0.5794,
        0.5794, 0.5794, 0.5794, 0.5794, 0.5794, 0.5794, 0.5794, 0.5794, 0.5794,
        0.5794, 0.5794, 0.5794, 0.5794, 0.5794, 0.3707, 0.1791]) return=
61352.37296508392
probs of actions:  tensor([0.0049, 0.2886, 0.5194, 0.5667, 0.0384, 0.5416,
0.2841, 0.5163, 0.5702,
        0.3191, 0.3114, 0.2961, 0.5570, 0.5329, 0.3011, 0.5474, 0.5238, 0.2829,
        0.0357, 0.5463, 0.2982, 0.5444, 0.3465, 0.6957, 0.9113],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4887, 0.4391, 0.3796, 0.3262, 0.2924, 0.2469, 0.2275,
0.2283, 0.2184,
        0.2066, 0.2082, 0.2093, 0.2147, 0.2084, 0.1992, 0.2071, 0.2027, 0.1950,
        0.1931, 0.2141, 0.2035, 0.2103, 0.2006, 0.2081, 0.2071])
finalReturns:  tensor([0.0365, 0.0446, 0.0280])
----------------------------------------
iter  0  stage  22  ep  59999   adversary:  const95-1.0,
```

```
  actions:  tensor([ 9, 12,  3,  9,  9,  9,  6, 12,  9,  6,  9,  3,  6, 12,  9,
9,  9,  6,
         9,  6,  9, 12,  9,  9,  0])
loss=  tensor(0.0352, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.6075,
0.6075, 0.6075, 0.6075, 0.6075, 0.6075, 0.6075, 0.6075, 0.6075,
        0.6075, 0.6075, 0.6075, 0.6075, 0.6075, 0.6075, 0.6075, 0.6075, 0.6075,
        0.6075, 0.6075, 0.6075, 0.6075, 0.6075, 0.3862, 0.1857]) return=
62546.458717917136
probs of actions:  tensor([0.5353, 0.0431, 0.0359, 0.5870, 0.6315, 0.5678,
0.2807, 0.0454, 0.5922,
        0.2558, 0.6168, 0.0411, 0.2766, 0.0482, 0.5834, 0.6188, 0.5877, 0.2788,
        0.5710, 0.2691, 0.5920, 0.0468, 0.7153, 0.6523, 0.9644],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4130, 0.3777, 0.3100, 0.2840, 0.2653, 0.2561,
0.2279, 0.2361,
        0.2347, 0.2186, 0.2243, 0.2066, 0.1916, 0.2085, 0.2095, 0.2104, 0.2155,
        0.2045, 0.2110, 0.2012, 0.1977, 0.2132, 0.2131, 0.2211])
finalReturns:  tensor([0.0399, 0.0480, 0.0355])
----------------------------------------
iter  0  stage  22  ep  89999   adversary:  const95-1.0,
  actions:  tensor([ 3,  9,  9,  9,  9,  9,  9,  9,  9,  9,  9,  9,  9, 15,
6,  9,  9,
        12,  9,  9,  9,  6,  9,  0])
loss=  tensor(0.0937, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.6155,
0.6155, 0.6155, 0.6155, 0.6155, 0.6155, 0.6155, 0.6155, 0.6155,
        0.6155, 0.6155, 0.6155, 0.6155, 0.6155, 0.6155, 0.6155, 0.6155, 0.6155,
        0.6155, 0.6155, 0.6155, 0.6155, 0.6155, 0.3906, 0.1875]) return=
63514.70077728213
probs of actions:  tensor([0.0296, 0.7851, 0.7652, 0.7515, 0.7909, 0.7315,
0.7606, 0.7464, 0.7608,
        0.7813, 0.7735, 0.7258, 0.7492, 0.7762, 0.0049, 0.1191, 0.7491, 0.7561,
        0.0529, 0.7435, 0.7497, 0.7477, 0.0676, 0.8504, 0.9806],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5103, 0.3999, 0.3478, 0.3110, 0.2848, 0.2658, 0.2520,
0.2419, 0.2345,
        0.2290, 0.2249, 0.2218, 0.2196, 0.2179, 0.2022, 0.2346, 0.2185, 0.2171,
        0.2097, 0.2223, 0.2199, 0.2181, 0.2213, 0.2088, 0.2179])
finalReturns:  tensor([0.0324, 0.0360, 0.0304])
----------------------------------------
iter  0  stage  22  ep  119999   adversary:  const95-1.0,
  actions:  tensor([ 9,  9,  9,  9, 12,  9,  6,  9,  9,  9,  9,  9,  9,  9,  9,
9,  9,  9,
         9,  9,  9,  3,  9,  9,  0])
loss=  tensor(0.0097, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.5764,
0.5764, 0.5764, 0.5764, 0.5764, 0.5764, 0.5764, 0.5764, 0.5764,
        0.5764, 0.5764, 0.5764, 0.5764, 0.5764, 0.5764, 0.5764, 0.5764, 0.5764,
        0.5764, 0.5764, 0.5764, 0.5764, 0.5764, 0.3690, 0.1784]) return=
63600.55185293454
```

```
probs of actions:  tensor([0.7547, 0.8389, 0.8221, 0.8118, 0.0509, 0.7931,
0.0775, 0.8037, 0.8180,
        0.8343, 0.8274, 0.7840, 0.8068, 0.8295, 0.8006, 0.8319, 0.8038, 0.8153,
        0.7895, 0.8013, 0.8043, 0.0150, 0.8904, 0.9033, 0.9875],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2854, 0.2788, 0.2660,
0.2413, 0.2340,
        0.2286, 0.2246, 0.2216, 0.2194, 0.2177, 0.2165, 0.2156, 0.2149, 0.2144,
        0.2140, 0.2137, 0.2135, 0.2205, 0.1993, 0.2026, 0.2132])
finalReturns:  tensor([0.0388, 0.0469, 0.0348])
----------------------------------------
iter  0  stage  22  ep  149999   adversary:  const95-1.0,
  actions:  tensor([ 9,  9,  9,  9, 12,  9,  9,  9,  9,  6, 12,  9,  9,  9,  0,
9,  9,  9,
         9,  9,  9,  9,  9,  9,  0])
loss=  tensor(0.0138, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.6015,
0.6015, 0.6015, 0.6015, 0.6015, 0.6015, 0.6015, 0.6015, 0.6015,
        0.6015, 0.6015, 0.6015, 0.6015, 0.6015, 0.6015, 0.6015, 0.6015, 0.6015,
        0.6015, 0.6015, 0.6015, 0.6015, 0.6015, 0.3829, 0.1843]) return=
63367.220552996965
probs of actions:  tensor([0.7048, 0.7939, 0.7766, 0.7712, 0.0921, 0.7481,
0.7808, 0.7570, 0.7766,
        0.0626, 0.0909, 0.7358, 0.7653, 0.7887, 0.0541, 0.7934, 0.7553, 0.7736,
        0.7408, 0.7540, 0.7546, 0.7588, 0.8251, 0.8820, 0.9942],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2854, 0.2788, 0.2615,
0.2489, 0.2396,
        0.2373, 0.2142, 0.2257, 0.2225, 0.2200, 0.2263, 0.1960, 0.2001, 0.2033,
        0.2056, 0.2074, 0.2088, 0.2098, 0.2105, 0.2111, 0.2196])
finalReturns:  tensor([0.0397, 0.0478, 0.0354])
----------------------------------------
iter  0  stage  21  ep  29999   adversary:  const95-1.0,
  actions:  tensor([ 9,  6,  9,  9, 12,  9, 12,  9,  9,  9,  9,  9,  9,  9,  6,
9,  6,  9,
         9, 12, 12,  9,  9,  9,  0])
loss=  tensor(0.0525, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.8064,
0.8064, 0.8064, 0.8064, 0.8064, 0.8064, 0.8064, 0.8064, 0.8064,
        0.8064, 0.8064, 0.8064, 0.8064, 0.8064, 0.8064, 0.8064, 0.8064, 0.8064,
        0.8064, 0.8064, 0.8064, 0.8064, 0.5759, 0.3687, 0.1783]) return=
63900.06666813643
probs of actions:  tensor([0.7450, 0.0332, 0.8032, 0.7765, 0.1419, 0.7829,
0.1524, 0.7768, 0.7725,
        0.7943, 0.7975, 0.7758, 0.7730, 0.7909, 0.0445, 0.7913, 0.0394, 0.7892,
        0.7642, 0.1584, 0.1688, 0.7797, 0.7790, 0.8710, 0.9861],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4238, 0.3523, 0.3142, 0.2808, 0.2754, 0.2527,
0.2547, 0.2439,
        0.2359, 0.2300, 0.2257, 0.2224, 0.2200, 0.2227, 0.2098, 0.2150, 0.2041,
```

```
        0.2063, 0.2016, 0.2099, 0.2225, 0.2200, 0.2182, 0.2250])
finalReturns:  tensor([0.0792, 0.0873, 0.0745, 0.0466])
----------------------------------------
iter  0  stage  21  ep  59999   adversary:  const95-1.0,
  actions:  tensor([ 9,  9,  9, 12,  9, 12,  9, 12,  9,  9,  9,  9, 12,  9,  9,
9,  3,  9,
         9, 12,  9,  9, 12,  9,  0])
loss=  tensor(0.1608, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.7875,
0.7875, 0.7875, 0.7875, 0.7875, 0.7875, 0.7875, 0.7875, 0.7875, 0.7875,
        0.7875, 0.7875, 0.7875, 0.7875, 0.7875, 0.7875, 0.7875, 0.7875, 0.7875,
        0.7875, 0.7875, 0.7875, 0.7875, 0.5642, 0.3623, 0.1756]) return=
64562.60586840481
probs of actions:  tensor([0.5976, 0.6168, 0.6466, 0.3209, 0.6422, 0.2996,
0.6183, 0.3199, 0.6098,
        0.6397, 0.6462, 0.6327, 0.3156, 0.6326, 0.6199, 0.6341, 0.0044, 0.6398,
        0.6093, 0.3274, 0.5959, 0.5767, 0.3889, 0.7362, 0.9915],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3143, 0.2999, 0.2705, 0.2679,
0.2472, 0.2506,
        0.2409, 0.2337, 0.2284, 0.2181, 0.2287, 0.2247, 0.2217, 0.2267, 0.2038,
        0.2060, 0.2014, 0.2160, 0.2152, 0.2083, 0.2213, 0.2272])
finalReturns:  tensor([0.0845, 0.0926, 0.0862, 0.0516])
----------------------------------------
iter  0  stage  21  ep  89999   adversary:  const95-1.0,
  actions:  tensor([ 9,  9, 12, 12, 12,  9, 12,  6,  9,  9,  9,  9,  9, 12, 12,
12, 12, 12,
         9,  9, 12, 12, 12,  9,  0])
loss=  tensor(0.1576, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.8257,
0.8257, 0.8257, 0.8257, 0.8257, 0.8257, 0.8257, 0.8257, 0.8257, 0.8257,
        0.8257, 0.8257, 0.8257, 0.8257, 0.8257, 0.8257, 0.8257, 0.8257, 0.8257,
        0.8257, 0.8257, 0.8257, 0.8257, 0.5877, 0.3752, 0.1811]) return=
65901.80171251376
probs of actions:  tensor([0.3774, 0.3514, 0.5611, 0.5640, 0.5644, 0.3992,
0.5829, 0.0208, 0.3723,
        0.4031, 0.4069, 0.4050, 0.3920, 0.5676, 0.5599, 0.5592, 0.5817, 0.5480,
        0.3768, 0.3752, 0.5988, 0.6690, 0.6565, 0.4812, 0.9943],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3550, 0.3230, 0.2999, 0.2895, 0.2629,
0.2668, 0.2419,
        0.2344, 0.2289, 0.2248, 0.2218, 0.2132, 0.2187, 0.2229, 0.2260, 0.2284,
        0.2365, 0.2305, 0.2197, 0.2236, 0.2266, 0.2351, 0.2375])
finalReturns:  tensor([0.0972, 0.1116, 0.0974, 0.0565])
----------------------------------------
iter  0  stage  21  ep  119999   adversary:  const95-1.0,
  actions:  tensor([12,  9,  9,  9, 12, 12, 12,  6, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0851, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.8567,
```

```
0.8567, 0.8567, 0.8567, 0.8567, 0.8567, 0.8567, 0.8567, 0.8567,
        0.8567, 0.8567, 0.8567, 0.8567, 0.8567, 0.8567, 0.8567, 0.8567, 0.8567,
        0.8567, 0.8567, 0.8567, 0.8567, 0.6066, 0.3857, 0.1854]) return=
67404.68062374518
probs of actions:  tensor([0.7116, 0.2010, 0.2402, 0.2483, 0.7271, 0.7151,
0.7406, 0.0122, 0.7393,
        0.7099, 0.7133, 0.7056, 0.7096, 0.7272, 0.7157, 0.7208, 0.7393, 0.7119,
        0.7350, 0.7298, 0.7515, 0.8319, 0.8060, 0.6590, 0.9953],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4292, 0.3682, 0.3255, 0.2889, 0.2751, 0.2649,
0.2682, 0.2366,
        0.2364, 0.2362, 0.2360, 0.2359, 0.2358, 0.2358, 0.2357, 0.2357, 0.2357,
        0.2357, 0.2356, 0.2356, 0.2356, 0.2356, 0.2356, 0.2500])
finalReturns:  tensor([0.1002, 0.1146, 0.0999, 0.0646])
----------------------------------------
iter  0  stage  21  ep  149999   adversary:  const95-1.0,
  actions:  tensor([12, 12,  9, 12, 12, 12, 12, 12,  9, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12,  9, 12, 12, 12,  0])
loss=  tensor(0.0369, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.8380,
0.8380, 0.8380, 0.8380, 0.8380, 0.8380, 0.8380, 0.8380, 0.8380,
        0.8380, 0.8380, 0.8380, 0.8380, 0.8380, 0.8380, 0.8380, 0.8380, 0.8380,
        0.8380, 0.8380, 0.8380, 0.8380, 0.5952, 0.3794, 0.1828]) return=
68049.04485884288
probs of actions:  tensor([0.8436, 0.8976, 0.1146, 0.8559, 0.8613, 0.8556,
0.8694, 0.8624, 0.1145,
        0.8490, 0.8528, 0.8457, 0.8438, 0.8603, 0.8480, 0.8552, 0.8680, 0.8512,
        0.8645, 0.8594, 0.1076, 0.9306, 0.9117, 0.8245, 0.9968],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3775, 0.3257, 0.3019, 0.2846, 0.2719,
0.2626, 0.2620,
        0.2430, 0.2411, 0.2397, 0.2387, 0.2379, 0.2373, 0.2369, 0.2366, 0.2363,
        0.2361, 0.2360, 0.2422, 0.2284, 0.2302, 0.2315, 0.2469])
finalReturns:  tensor([0.0990, 0.1134, 0.0991, 0.0641])
----------------------------------------
iter  0  stage  21  ep  179999   adversary:  const95-1.0,
  actions:  tensor([12, 12, 12, 12, 12,  9, 12, 12, 12, 12,  9, 12,  9, 12, 12,
12, 12,  9,
        12, 12,  9, 12, 12, 12,  0])
loss=  tensor(0.0208, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.8275,
0.8275, 0.8275, 0.8275, 0.8275, 0.8275, 0.8275, 0.8275, 0.8275,
        0.8275, 0.8275, 0.8275, 0.8275, 0.8275, 0.8275, 0.8275, 0.8275, 0.8275,
        0.8275, 0.8275, 0.8275, 0.8275, 0.5888, 0.3759, 0.1813]) return=
67652.69436488167
probs of actions:  tensor([0.9003, 0.9414, 0.9204, 0.9110, 0.9162, 0.0734,
0.9213, 0.9156, 0.9173,
        0.9058, 0.0760, 0.9050, 0.0810, 0.9148, 0.9034, 0.9105, 0.9194, 0.0766,
        0.9171, 0.9125, 0.0652, 0.9634, 0.9505, 0.8924, 0.9977],
```

```
       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, 0.2517, 0.2354, 0.2418, 0.2280, 0.2299, 0.2313, 0.2324, 0.2395,
        0.2264, 0.2287, 0.2367, 0.2243, 0.2271, 0.2292, 0.2452])
finalReturns:  tensor([0.0983, 0.1127, 0.0986, 0.0639])
----------------------------------------
iter  0  stage  20  ep  29999   adversary:  const95-1.0,
  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(0.0203, 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([0.9466, 0.9715, 0.9604, 0.9532, 0.9564, 0.9554,
0.9586, 0.9567, 0.9564,
        0.9497, 0.9536, 0.9498, 0.9472, 0.9553, 0.9484, 0.9530, 0.9583, 0.9523,
        0.9568, 0.9543, 0.9625, 0.9825, 0.9769, 0.9455, 0.9969],
       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  20  ep  59999   adversary:  const95-1.0,
  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(0.0137, 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([0.9620, 0.9807, 0.9729, 0.9670, 0.9694, 0.9691,
0.9710, 0.9699, 0.9693,
        0.9641, 0.9674, 0.9646, 0.9624, 0.9685, 0.9634, 0.9670, 0.9707, 0.9663,
        0.9698, 0.9678, 0.9753, 0.9893, 0.9836, 0.9615, 0.9975],
       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  20  ep  89999   adversary:  const95-1.0,
```

```
  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(0.0081, 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([0.9754, 0.9882, 0.9832, 0.9790, 0.9807, 0.9806,
0.9816, 0.9810, 0.9803,
        0.9769, 0.9792, 0.9774, 0.9758, 0.9799, 0.9765, 0.9789, 0.9813, 0.9785,
        0.9809, 0.9793, 0.9857, 0.9934, 0.9898, 0.9773, 0.9982],
       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  20  ep  119999   adversary:  const95-1.0,
  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(0.0072, 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([0.9778, 0.9895, 0.9850, 0.9812, 0.9826, 0.9827,
0.9834, 0.9828, 0.9822,
        0.9793, 0.9814, 0.9798, 0.9781, 0.9820, 0.9790, 0.9810, 0.9832, 0.9807,
        0.9829, 0.9813, 0.9868, 0.9942, 0.9910, 0.9807, 0.9991],
       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  20  ep  149999   adversary:  const95-1.0,
  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(0.0048, 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([0.9843, 0.9929, 0.9897, 0.9868, 0.9879, 0.9880,
0.9884, 0.9880, 0.9875,
        0.9853, 0.9870, 0.9858, 0.9845, 0.9873, 0.9851, 0.9867, 0.9882, 0.9864,
        0.9880, 0.9869, 0.9912, 0.9964, 0.9940, 0.9866, 0.9994],
       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  20  ep  179999   adversary:  const95-1.0,
  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(0.0030, 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([0.9895, 0.9955, 0.9934, 0.9913, 0.9921, 0.9922,
0.9924, 0.9922, 0.9918,
        0.9902, 0.9915, 0.9906, 0.9896, 0.9916, 0.9901, 0.9913, 0.9923, 0.9911,
        0.9922, 0.9914, 0.9946, 0.9980, 0.9964, 0.9908, 0.9998],
       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  20  ep  209999   adversary:  const95-1.0,
  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(0.0035, 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([0.9880, 0.9947, 0.9924, 0.9900, 0.9908, 0.9909,
0.9912, 0.9910, 0.9905,
        0.9885, 0.9903, 0.9891, 0.9878, 0.9903, 0.9884, 0.9900, 0.9912, 0.9897,
        0.9910, 0.9901, 0.9938, 0.9978, 0.9962, 0.9878, 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  20  ep  239999   adversary:  const95-1.0,
  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(0.0035, 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, 1.0332, 0.7825, 0.5612, 0.3606, 0.1749]) return=
68694.09895647192
probs of actions:  tensor([0.9881, 0.9948, 0.9924, 0.9900, 0.9908, 0.9911,
0.9913, 0.9911, 0.9906,
        0.9887, 0.9903, 0.9892, 0.9880, 0.9903, 0.9886, 0.9900, 0.9913, 0.9897,
        0.9911, 0.9902, 0.9942, 0.9978, 0.9961, 0.9881, 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  29999   adversary:  const95-1.0,
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12,  9, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0048, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.1986,
1.1986, 1.1986, 1.1986, 1.1986, 1.1986, 1.1986, 1.1986, 1.1986, 1.1986,
        1.1986, 1.1986, 1.1986, 1.1986, 1.1986, 1.1986, 1.1986, 1.1986, 1.1986,
        1.1986, 1.1986, 0.9485, 0.7275, 0.5271, 0.3416, 0.1669]) return=
68460.61730250808
probs of actions:  tensor([0.9910, 0.9962, 0.9944, 0.9924, 0.9931, 0.9932,
0.9934, 0.9933, 0.9929,
        0.9913, 0.0065, 0.9918, 0.9906, 0.9926, 0.9912, 0.9925, 0.9934, 0.9922,
        0.9933, 0.9931, 0.9957, 0.9986, 0.9976, 0.9906, 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, 0.2542, 0.2372, 0.2368, 0.2365, 0.2363, 0.2361, 0.2360, 0.2359,
        0.2358, 0.2358, 0.2357, 0.2357, 0.2357, 0.2357, 0.2500])
finalReturns:  tensor([0.2299, 0.2443, 0.2296, 0.1942, 0.1441, 0.0832])
----------------------------------------
iter  0  stage  19  ep  59999   adversary:  const95-1.0,
  actions:  tensor([12, 12, 12, 12, 12, 12, 12,  9, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0037, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.1999,
```

```
1.1999, 1.1999, 1.1999, 1.1999, 1.1999, 1.1999, 1.1999, 1.1999,
        1.1999, 1.1999, 1.1999, 1.1999, 1.1999, 1.1999, 1.1999, 1.1999, 1.1999,
        1.1999, 1.1999, 0.9493, 0.7280, 0.5275, 0.3418, 0.1670]) return=
68453.26034712071
probs of actions:  tensor([0.9926, 0.9969, 0.9955, 0.9938, 0.9944, 0.9945,
0.9947, 0.0049, 0.9943,
        0.9929, 0.9941, 0.9932, 0.9924, 0.9940, 0.9928, 0.9939, 0.9947, 0.9937,
        0.9946, 0.9948, 0.9966, 0.9989, 0.9982, 0.9924, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3345, 0.3083, 0.2892, 0.2753,
0.2714, 0.2498,
        0.2462, 0.2436, 0.2416, 0.2401, 0.2389, 0.2381, 0.2375, 0.2370, 0.2367,
        0.2364, 0.2362, 0.2360, 0.2359, 0.2358, 0.2358, 0.2501])
finalReturns:  tensor([0.2300, 0.2444, 0.2297, 0.1943, 0.1441, 0.0832])
----------------------------------------
iter  0  stage  19  ep  89999   adversary:  const95-1.0,
  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(0.0027, 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, 0.9500, 0.7285, 0.5277, 0.3420, 0.1670]) return=
68694.09895647192
probs of actions:  tensor([0.9943, 0.9977, 0.9966, 0.9953, 0.9958, 0.9958,
0.9960, 0.9959, 0.9956,
        0.9946, 0.9956, 0.9948, 0.9942, 0.9954, 0.9945, 0.9954, 0.9960, 0.9952,
        0.9959, 0.9963, 0.9976, 0.9993, 0.9987, 0.9942, 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  19  ep  119999   adversary:  const95-1.0,
  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(0.0022, 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, 0.9500, 0.7285, 0.5277, 0.3420, 0.1670]) return=
68694.09895647192
probs of actions:  tensor([0.9951, 0.9981, 0.9971, 0.9959, 0.9964, 0.9964,
0.9965, 0.9965, 0.9963,
        0.9953, 0.9962, 0.9956, 0.9949, 0.9961, 0.9952, 0.9961, 0.9966, 0.9959,
        0.9965, 0.9970, 0.9980, 0.9995, 0.9990, 0.9948, 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  19  ep  149999   adversary:  const95-1.0,
  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([1.2009,
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([0.9956, 0.9982, 0.9974, 0.9963, 0.9966, 0.9967,
0.9968, 0.9969, 0.9966,
        0.9957, 0.9966, 0.9959, 0.9954, 0.9963, 0.9956, 0.9964, 0.9969, 0.9963,
        0.9968, 0.9974, 0.9982, 0.9997, 0.9993, 0.9946, 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  19  ep  179999   adversary:  const95-1.0,
  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(0.0021, 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, 0.9500, 0.7285, 0.5277, 0.3420, 0.1670]) return=
68694.09895647192
probs of actions:  tensor([0.9953, 0.9981, 0.9972, 0.9960, 0.9964, 0.9965,
0.9966, 0.9967, 0.9964,
        0.9954, 0.9964, 0.9956, 0.9951, 0.9961, 0.9953, 0.9962, 0.9967, 0.9960,
        0.9966, 0.9973, 0.9982, 0.9997, 0.9992, 0.9942, 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  19  ep  209999   adversary:  const95-1.0,
```

```
  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(0.0016, 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, 0.9500, 0.7285, 0.5277, 0.3420, 0.1670]) return=
68694.09895647192
probs of actions:  tensor([0.9963, 0.9985, 0.9978, 0.9969, 0.9972, 0.9973,
0.9974, 0.9974, 0.9972,
        0.9964, 0.9972, 0.9966, 0.9961, 0.9969, 0.9963, 0.9970, 0.9974, 0.9969,
        0.9973, 0.9979, 0.9987, 1.0000, 0.9994, 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([0.2301, 0.2445, 0.2297, 0.1944, 0.1441, 0.0832])
----------------------------------------
iter  0  stage  19  ep  239999   adversary:  const95-1.0,
  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(0.0015, 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, 0.9500, 0.7285, 0.5277, 0.3420, 0.1670]) return=
68694.09895647192
probs of actions:  tensor([0.9963, 0.9985, 0.9978, 0.9969, 0.9972, 0.9973,
0.9973, 0.9974, 0.9971,
        0.9964, 0.9972, 0.9966, 0.9960, 0.9969, 0.9962, 0.9970, 0.9974, 0.9968,
        0.9973, 0.9978, 0.9986, 1.0000, 0.9995, 0.9958, 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  19  ep  269999   adversary:  const95-1.0,
  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(0.0010, 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, 0.9500, 0.7285, 0.5277, 0.3420, 0.1670]) return=
68694.09895647192
```

```
probs of actions:  tensor([0.9975, 0.9990, 0.9986, 0.9979, 0.9981, 0.9982,
0.9982, 0.9983, 0.9981,
        0.9975, 0.9981, 0.9977, 0.9973, 0.9979, 0.9975, 0.9980, 0.9983, 0.9979,
        0.9982, 0.9987, 0.9993, 1.0000, 0.9997, 0.9971, 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  19  ep  280092   adversary:  const95-1.0,
  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([1.2009,
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([0.9981, 0.9993, 0.9989, 0.9984, 0.9986, 0.9987,
0.9987, 0.9987, 0.9985,
        0.9981, 0.9986, 0.9982, 0.9979, 0.9984, 0.9981, 0.9985, 0.9987, 0.9984,
        0.9987, 0.9990, 0.9995, 1.0000, 0.9998, 0.9979, 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  607   adversary:  const95-1.0,
  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(0.0011, 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, 1.3631, 1.1119, 0.8901, 0.6892, 0.5034, 0.3283, 0.1613]) return=
68694.09895647192
probs of actions:  tensor([0.9982, 0.9994, 0.9990, 0.9985, 0.9987, 0.9988,
0.9988, 0.9988, 0.9986,
        0.9983, 0.9987, 0.9984, 0.9981, 0.9986, 0.9982, 0.9986, 0.9988, 0.9985,
        0.9990, 0.9991, 0.9995, 1.0000, 0.9998, 0.9981, 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  4304   adversary:  const95-1.0,
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12,  9, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0015, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.4969,
1.4969, 1.4969, 1.4969, 1.4969, 1.4969, 1.4969, 1.4969, 1.4969,
        1.4969, 1.4969, 1.4969, 1.4969, 1.4969, 1.4969, 1.4969, 1.4969, 1.4969,
        1.2527, 1.0359, 0.8386, 0.6553, 0.4822, 0.3165, 0.1562]) return=
68487.85710629346
probs of actions:  tensor([0.9984, 0.9994, 0.9991, 0.9987, 0.9989, 0.9989,
0.9989, 0.9990, 0.9988,
        0.9985, 0.9989, 0.9986, 0.9983, 0.9988, 0.9984, 0.9988, 0.0010, 0.9990,
        0.9992, 0.9992, 0.9996, 1.0000, 0.9999, 0.9983, 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, 0.2441, 0.2298,
        0.2312, 0.2323, 0.2331, 0.2337, 0.2342, 0.2346, 0.2492])
finalReturns:  tensor([0.3812, 0.3956, 0.3812, 0.3462, 0.2964, 0.2358, 0.1673,
0.0930])
----------------------------------------
iter  0  stage  16  ep  30   adversary:  const95-1.0,
  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(0.0021, 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, 1.4248,
        1.2024, 1.0011, 0.8148, 0.6396, 0.4723, 0.3109, 0.1538]) return=
68694.09895647192
probs of actions:  tensor([0.9984, 0.9995, 0.9991, 0.9987, 0.9989, 0.9989,
0.9989, 0.9990, 0.9988,
        0.9985, 0.9989, 0.9986, 0.9983, 0.9988, 0.9984, 0.9988, 0.9990, 0.9990,
        0.9992, 0.9992, 0.9996, 1.0000, 0.9999, 0.9983, 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  10043   adversary:  const95-1.0,
  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(0.0026, 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, 1.5780, 1.3551,
        1.1534, 0.9669, 0.7914, 0.6240, 0.4625, 0.3054, 0.1515]) return=
68694.09895647192
probs of actions:  tensor([0.9986, 0.9995, 0.9992, 0.9989, 0.9990, 0.9990,
0.9990, 0.9991, 0.9989,
        0.9986, 0.9990, 0.9987, 0.9985, 0.9989, 0.9986, 0.9990, 0.9992, 0.9991,
        0.9993, 0.9993, 0.9997, 1.0000, 0.9999, 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([0.5504, 0.5648, 0.5499, 0.5144, 0.4641, 0.4030, 0.3342,
0.2596, 0.1807,
        0.0987])
----------------------------------------
iter  0  stage  14  ep  968   adversary:  const95-1.0,
  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(0.0032, 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([0.9987, 0.9995, 0.9993, 0.9989, 0.9991, 0.9991,
0.9991, 0.9991, 0.9990,
        0.9987, 0.9991, 0.9988, 0.9986, 0.9990, 0.9990, 0.9991, 0.9993, 0.9992,
        0.9994, 0.9994, 0.9997, 1.0000, 0.9999, 0.9986, 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  36   adversary:  const95-1.0,
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,  9,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(4.9667, 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=
68474.29269380601
probs of actions:  tensor([9.9870e-01, 9.9955e-01, 9.9930e-01, 9.9896e-01,
9.9907e-01, 9.9912e-01,
        9.9911e-01, 9.9914e-01, 9.9902e-01, 9.9874e-01, 9.9907e-01, 9.9884e-01,
        9.9861e-01, 9.9900e-01, 9.8588e-04, 9.9909e-01, 9.9929e-01, 9.9916e-01,
        9.9939e-01, 9.9939e-01, 9.9970e-01, 1.0000e+00, 9.9995e-01, 9.9856e-01,
        1.0000e+00], 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, 0.2457, 0.2310, 0.2321, 0.2330,
        0.2337, 0.2341, 0.2345, 0.2348, 0.2350, 0.2351, 0.2497])
finalReturns:  tensor([0.7026, 0.7170, 0.6958, 0.6677, 0.6228, 0.5659, 0.5001,
0.4277, 0.3506,
        0.2699, 0.1865, 0.1012])
----------------------------------------
iter  0  stage  12  ep  7589   adversary:  const95-1.0,
  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(0.0042, 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([0.9990, 0.9996, 0.9994, 0.9992, 0.9993, 0.9993,
0.9993, 0.9993, 0.9992,
        0.9990, 0.9993, 0.9991, 0.9990, 0.9992, 0.9992, 0.9993, 0.9995, 0.9993,
        0.9996, 0.9996, 0.9998, 1.0000, 1.0000, 0.9989, 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  0   adversary:  const95-1.0,
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,  9, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(6.1888, 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=
68466.13679373146
probs of actions:  tensor([9.9895e-01, 9.9965e-01, 9.9944e-01, 9.9917e-01,
```

```
9.9926e-01, 9.9930e-01,
        9.9929e-01, 9.9931e-01, 9.9921e-01, 9.9899e-01, 9.9926e-01, 9.9907e-01,
        9.9921e-04, 9.9925e-01, 9.9920e-01, 9.9927e-01, 9.9947e-01, 9.9932e-01,
        9.9959e-01, 9.9959e-01, 9.9979e-01, 1.0000e+00, 9.9997e-01, 9.9888e-01,
        1.0000e+00], 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, 0.2488, 0.2332, 0.2338, 0.2343, 0.2346, 0.2348,
        0.2350, 0.2352, 0.2353, 0.2354, 0.2354, 0.2355, 0.2499])
finalReturns:  tensor([0.8808, 0.8952, 0.8738, 0.8454, 0.8004, 0.7432, 0.6772,
0.6048, 0.5275,
        0.4468, 0.3633, 0.2779, 0.1911, 0.1032])
----------------------------------------
iter  0  stage  10  ep  0   adversary:  const95-1.0,
  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(0.0064, 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([0.9990, 0.9996, 0.9994, 0.9992, 0.9993, 0.9993,
0.9993, 0.9993, 0.9992,
        0.9990, 0.9993, 0.9991, 0.9990, 0.9992, 0.9992, 0.9993, 0.9995, 0.9993,
        0.9996, 0.9996, 0.9998, 1.0000, 1.0000, 0.9989, 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  22   adversary:  const95-1.0,
  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(0.0078, 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([0.9989, 0.9996, 0.9994, 0.9992, 0.9992, 0.9993,
0.9993, 0.9993, 0.9992,
        0.9990, 0.9993, 0.9991, 0.9990, 0.9992, 0.9992, 0.9993, 0.9995, 0.9993,
        0.9996, 0.9996, 0.9998, 1.0000, 1.0000, 0.9989, 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:  const95-1.0,
  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(0.0091, 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([0.9989, 0.9996, 0.9994, 0.9992, 0.9992, 0.9993,
0.9993, 0.9993, 0.9992,
        0.9990, 0.9993, 0.9991, 0.9990, 0.9992, 0.9992, 0.9993, 0.9995, 0.9993,
        0.9996, 0.9996, 0.9998, 1.0000, 1.0000, 0.9989, 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:  const95-1.0,
  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(0.0103, 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([0.9989, 0.9996, 0.9994, 0.9992, 0.9992, 0.9993,
0.9993, 0.9993, 0.9992,
        0.9990, 0.9993, 0.9991, 0.9990, 0.9992, 0.9992, 0.9993, 0.9995, 0.9993,
        0.9996, 0.9996, 0.9998, 1.0000, 1.0000, 0.9989, 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:  const95-1.0,
  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(0.0116, 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([0.9989, 0.9996, 0.9994, 0.9992, 0.9992, 0.9993,
0.9993, 0.9993, 0.9992,
        0.9990, 0.9993, 0.9991, 0.9990, 0.9992, 0.9992, 0.9993, 0.9995, 0.9993,
        0.9996, 0.9996, 0.9998, 1.0000, 1.0000, 0.9989, 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:  const95-1.0,
  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(0.0129, 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([0.9989, 0.9996, 0.9994, 0.9992, 0.9992, 0.9993,
0.9993, 0.9993, 0.9992,
        0.9990, 0.9993, 0.9991, 0.9990, 0.9992, 0.9992, 0.9993, 0.9995, 0.9993,
        0.9996, 0.9996, 0.9998, 1.0000, 1.0000, 0.9989, 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:  const95-1.0,
  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(0.0144, 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([0.9989, 0.9996, 0.9994, 0.9992, 0.9992, 0.9993,
0.9993, 0.9993, 0.9992,
        0.9990, 0.9993, 0.9991, 0.9990, 0.9992, 0.9992, 0.9993, 0.9995, 0.9993,
        0.9996, 0.9996, 0.9998, 1.0000, 1.0000, 0.9989, 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:  const95-1.0,
  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(0.0161, 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([0.9989, 0.9996, 0.9994, 0.9992, 0.9992, 0.9993,
0.9993, 0.9993, 0.9992,
        0.9990, 0.9993, 0.9991, 0.9990, 0.9992, 0.9992, 0.9993, 0.9995, 0.9993,
        0.9996, 0.9996, 0.9998, 1.0000, 1.0000, 0.9989, 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:  const95-1.0,
  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(0.0174, 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([0.9989, 0.9996, 0.9994, 0.9992, 0.9992, 0.9993,
0.9993, 0.9993, 0.9992,
        0.9990, 0.9993, 0.9991, 0.9990, 0.9992, 0.9992, 0.9993, 0.9995, 0.9993,
        0.9996, 0.9996, 0.9998, 1.0000, 1.0000, 0.9989, 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:  const95-1.0,
  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(0.0184, 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([0.9989, 0.9996, 0.9994, 0.9992, 0.9992, 0.9993,
0.9993, 0.9993, 0.9992,
        0.9990, 0.9993, 0.9991, 0.9990, 0.9992, 0.9992, 0.9993, 0.9995, 0.9993,
        0.9996, 0.9996, 0.9998, 1.0000, 1.0000, 0.9989, 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  88   adversary:  const95-1.0,
```

```
  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(0.0200, 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([0.9990, 0.9997, 0.9995, 0.9992, 0.9993, 0.9993,
0.9993, 0.9993, 0.9992,
        0.9991, 0.9993, 0.9991, 0.9990, 0.9992, 0.9992, 0.9993, 0.9995, 0.9993,
        0.9996, 0.9996, 0.9998, 1.0000, 1.0000, 0.9989, 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])
policy saved!
1,3,[5e-06,1][1, 10000, 1, 1],1683165834 saved
policy reset
----------------------------------------
iter  1  stage  24  ep  29999   adversary:  const95-1.0,
  actions:  tensor([42,  3,  3, 45, 18,  0,  3,  0,  3, 24,  3,  9,  3,  3,  0,
6,  3, 12,
         3,  6,  3, 45, 12,  3,  6])
loss=  tensor(-0.0060, grad_fn=<NegBackward0>)   ,  base rewards=
tensor([0.2448, 0.2448, 0.2448, 0.2448, 0.2448, 0.2448, 0.2448, 0.2448, 0.2448,
        0.2448, 0.2448, 0.2448, 0.2448, 0.2448, 0.2448, 0.2448, 0.2448, 0.2448,
        0.2448, 0.2448, 0.2448, 0.2448, 0.2448, 0.2448, 0.2448]) return=
62536.791981005576
probs of actions:  tensor([0.0082, 0.2535, 0.2494, 0.0080, 0.0216, 0.2569,
0.2423, 0.2699, 0.2563,
        0.0137, 0.2403, 0.1029, 0.2541, 0.2374, 0.2605, 0.1688, 0.2393, 0.0449,
        0.2451, 0.1836, 0.2257, 0.0110, 0.0487, 0.2466, 0.1903],
       grad_fn=<ExpBackward0>)
rewards:  tensor([ 0.3348,  0.5412,  0.4277,  0.1498,  0.3935,  0.3962,  0.3207,
0.2786,
         0.2401,  0.1639,  0.2572,  0.2257,  0.2296,  0.2131,  0.2020,  0.1831,
         0.1875,  0.1688,  0.1980,  0.1874,  0.1907, -0.0169,  0.2672,  0.2726,
         0.2412])
finalReturns:  tensor([-0.0036])
----------------------------------------
iter  1  stage  23  ep  29999   adversary:  const95-1.0,
  actions:  tensor([ 3,  0,  0, 24,  3,  0,  6,  0,  3,  9,  0,  3,  3,  3,  3,
```

```
3,  3,  6,
         3,  3,  3,  0,  3,  6,  3])
loss=  tensor(0.0199, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.3255,
0.3255, 0.3255, 0.3255, 0.3255, 0.3255, 0.3255, 0.3255, 0.3255,
        0.3255, 0.3255, 0.3255, 0.3255, 0.3255, 0.3255, 0.3255, 0.3255, 0.3255,
        0.3255, 0.3255, 0.3255, 0.3255, 0.3255, 0.3255, 0.1601]) return=
54781.59028774685
probs of actions:  tensor([0.4734, 0.2537, 0.2662, 0.0015, 0.4600, 0.2560,
0.1655, 0.2745, 0.4745,
        0.0526, 0.2590, 0.4529, 0.4598, 0.4480, 0.4363, 0.4599, 0.4514, 0.1761,
        0.4566, 0.4411, 0.4323, 0.2937, 0.4565, 0.2538, 0.4391],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5103, 0.4080, 0.3295, 0.2186, 0.3007, 0.2646, 0.2275,
0.2214, 0.1997,
        0.1841, 0.1992, 0.1838, 0.1796, 0.1764, 0.1741, 0.1724, 0.1711, 0.1674,
        0.1756, 0.1735, 0.1719, 0.1716, 0.1637, 0.1619, 0.1714])
finalReturns:  tensor([0.0077, 0.0113])
----------------------------------------
iter  1  stage  23  ep  59999   adversary:  const95-1.0,
  actions:  tensor([ 0,  6,  6,  3,  3,  6,  6,  6,  6,  9,  3,  0, 12,  3,  6,
3,  0,  6,
         6,  0,  3,  9,  0,  6,  0])
loss=  tensor(0.0099, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.3391,
0.3391, 0.3391, 0.3391, 0.3391, 0.3391, 0.3391, 0.3391, 0.3391,
        0.3391, 0.3391, 0.3391, 0.3391, 0.3391, 0.3391, 0.3391, 0.3391, 0.3391,
        0.3391, 0.3391, 0.3391, 0.3391, 0.3391, 0.3391, 0.1658]) return=
56244.87614304111
probs of actions:  tensor([0.1596, 0.4525, 0.4438, 0.3103, 0.3243, 0.4194,
0.3663, 0.4225, 0.4123,
        0.0672, 0.3033, 0.1854, 0.0129, 0.2958, 0.4512, 0.3226, 0.1837, 0.3981,
        0.3993, 0.1956, 0.2979, 0.0727, 0.1676, 0.7251, 0.5688],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5112, 0.3949, 0.3368, 0.2989, 0.2625, 0.2339, 0.2225,
0.2142, 0.2080,
        0.1989, 0.2095, 0.1994, 0.1704, 0.1993, 0.1883, 0.1914, 0.1861, 0.1716,
        0.1761, 0.1831, 0.1721, 0.1637, 0.1835, 0.1697, 0.1783])
finalReturns:  tensor([0.0088, 0.0124])
----------------------------------------
iter  1  stage  23  ep  89999   adversary:  const95-1.0,
  actions:  tensor([3, 6, 6, 6, 6, 0, 0, 6, 6, 3, 3, 0, 3, 3, 3, 9, 6, 0, 6, 6,
6, 6, 6, 6,
        0])
loss=  tensor(0.0037, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.3663,
0.3663, 0.3663, 0.3663, 0.3663, 0.3663, 0.3663, 0.3663, 0.3663,
        0.3663, 0.3663, 0.3663, 0.3663, 0.3663, 0.3663, 0.3663, 0.3663, 0.3663,
        0.3663, 0.3663, 0.3663, 0.3663, 0.3663, 0.3663, 0.1773]) return=
55808.994116578215
probs of actions:  tensor([0.1982, 0.5378, 0.5276, 0.4819, 0.4884, 0.1662,
```

```
0.2449, 0.5050, 0.4959,
        0.2201, 0.1924, 0.2021, 0.2375, 0.1857, 0.1872, 0.1046, 0.4413, 0.2141,
        0.4682, 0.4636, 0.5007, 0.4585, 0.4863, 0.8265, 0.8623],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5103, 0.4044, 0.3434, 0.3009, 0.2708, 0.2529, 0.2229,
0.1981, 0.1961,
        0.1972, 0.1895, 0.1847, 0.1733, 0.1717, 0.1706, 0.1625, 0.1790, 0.1853,
        0.1710, 0.1757, 0.1792, 0.1819, 0.1839, 0.1854, 0.1901])
finalReturns:  tensor([0.0093, 0.0129])
----------------------------------------
iter  1  stage  22  ep  29999   adversary:  const95-1.0,
  actions:  tensor([ 6,  9,  6,  6,  6,  9,  0,  9,  6,  9,  9,  9,  3,  6,  6,
9,  0, 12,
         6,  6,  6,  9,  6,  6,  0])
loss=  tensor(0.0535, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.5674,
0.5674, 0.5674, 0.5674, 0.5674, 0.5674, 0.5674, 0.5674, 0.5674, 0.5674,
        0.5674, 0.5674, 0.5674, 0.5674, 0.5674, 0.5674, 0.5674, 0.5674, 0.5674,
        0.5674, 0.5674, 0.5674, 0.5674, 0.5674, 0.3640, 0.1763]) return=
59812.40403986201
probs of actions:  tensor([0.4401, 0.4532, 0.4122, 0.4032, 0.3970, 0.4278,
0.0902, 0.4619, 0.3996,
        0.4055, 0.4287, 0.4020, 0.0991, 0.4210, 0.4286, 0.4193, 0.0713, 0.0278,
        0.3751, 0.3976, 0.3955, 0.4142, 0.3706, 0.4904, 0.9183],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5076, 0.4095, 0.3590, 0.3118, 0.2786, 0.2504, 0.2488,
0.2120, 0.2167,
        0.2053, 0.2072, 0.2086, 0.2168, 0.2011, 0.1983, 0.1917, 0.2050, 0.1745,
        0.1997, 0.1973, 0.1954, 0.1896, 0.1998, 0.1973, 0.1991])
finalReturns:  tensor([0.0288, 0.0324, 0.0227])
----------------------------------------
iter  1  stage  22  ep  59999   adversary:  const95-1.0,
  actions:  tensor([9, 9, 9, 9, 9, 6, 9, 0, 9, 6, 6, 0, 9, 6, 9, 9, 9, 9, 0, 9,
9, 6, 9, 9,
        9])
loss=  tensor(0.1111, 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, 0.3634, 0.1761]) return=
60649.09398743877
probs of actions:  tensor([0.6854, 0.7487, 0.7432, 0.7094, 0.7146, 0.1807,
0.6713, 0.0347, 0.7256,
        0.1790, 0.1936, 0.0563, 0.6536, 0.1959, 0.6958, 0.6811, 0.6686, 0.6919,
        0.0458, 0.6781, 0.6969, 0.1834, 0.8403, 0.8036, 0.0253],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2753, 0.2480,
0.2471, 0.2107,
        0.2158, 0.2092, 0.2079, 0.1828, 0.1946, 0.1890, 0.1948, 0.1992, 0.2026,
        0.2132, 0.1867, 0.1930, 0.2024, 0.1948, 0.1992, 0.2026])
```

```
finalReturns:  tensor([0.0303, 0.0384, 0.0265])
----------------------------------------
iter  1  stage  22  ep  89999   adversary:  const95-1.0,
  actions:  tensor([9, 9, 0, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 6, 9, 6, 9, 9, 9,
9, 9, 9, 9,
        0])
loss=  tensor(0.0092, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.6015,
0.6015, 0.6015, 0.6015, 0.6015, 0.6015, 0.6015, 0.6015, 0.6015,
        0.6015, 0.6015, 0.6015, 0.6015, 0.6015, 0.6015, 0.6015, 0.6015, 0.6015,
        0.6015, 0.6015, 0.6015, 0.6015, 0.6015, 0.3829, 0.1843]) return=
62572.5693677297
probs of actions:  tensor([0.7904, 0.8445, 0.0282, 0.8092, 0.8090, 0.8079,
0.7612, 0.8194, 0.8199,
        0.7612, 0.7964, 0.7755, 0.7516, 0.7987, 0.1141, 0.7796, 0.1087, 0.7876,
        0.8168, 0.7721, 0.7937, 0.7671, 0.9175, 0.9006, 0.9779],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3694, 0.2953, 0.2735, 0.2576, 0.2460,
0.2375, 0.2312,
        0.2265, 0.2231, 0.2205, 0.2185, 0.2171, 0.2205, 0.2082, 0.2138, 0.2032,
        0.2056, 0.2074, 0.2087, 0.2098, 0.2105, 0.2111, 0.2196])
finalReturns:  tensor([0.0397, 0.0478, 0.0354])
----------------------------------------
iter  1  stage  22  ep  119999   adversary:  const95-1.0,
  actions:  tensor([9, 9, 3, 9, 9, 9, 9, 6, 9, 9, 6, 9, 9, 9, 9, 9, 6, 9, 9, 9,
0, 9, 9, 9,
        0])
loss=  tensor(0.0061, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.5680,
0.5680, 0.5680, 0.5680, 0.5680, 0.5680, 0.5680, 0.5680, 0.5680,
        0.5680, 0.5680, 0.5680, 0.5680, 0.5680, 0.5680, 0.5680, 0.5680, 0.5680,
        0.5680, 0.5680, 0.5680, 0.5680, 0.5680, 0.3644, 0.1765]) return=
62135.91864110961
probs of actions:  tensor([0.8372, 0.8838, 0.0120, 0.8512, 0.8532, 0.8522,
0.8053, 0.0686, 0.8631,
        0.8069, 0.0854, 0.8223, 0.7987, 0.8417, 0.8412, 0.8223, 0.0815, 0.8315,
        0.8569, 0.8163, 0.0252, 0.8151, 0.9453, 0.9285, 0.9871],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3685, 0.3037, 0.2795, 0.2620, 0.2492,
0.2444, 0.2256,
        0.2224, 0.2245, 0.2111, 0.2115, 0.2118, 0.2121, 0.2123, 0.2169, 0.2055,
        0.2073, 0.2087, 0.2178, 0.1899, 0.1955, 0.1998, 0.2111])
finalReturns:  tensor([0.0384, 0.0465, 0.0346])
----------------------------------------
iter  1  stage  22  ep  149999   adversary:  const95-1.0,
  actions:  tensor([9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 3, 9, 9, 6, 9, 9, 9, 9,
9, 9, 9, 9,
        0])
loss=  tensor(0.0045, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.6024,
0.6024, 0.6024, 0.6024, 0.6024, 0.6024, 0.6024, 0.6024, 0.6024,
```

```
        0.6024, 0.6024, 0.6024, 0.6024, 0.6024, 0.6024, 0.6024, 0.6024, 0.6024,
        0.6024, 0.6024, 0.6024, 0.6024, 0.6024, 0.3834, 0.1845]) return=
63171.85534316236
probs of actions:  tensor([0.8761, 0.9140, 0.9068, 0.8864, 0.8894, 0.8893,
0.8472, 0.8918, 0.8966,
        0.8503, 0.8758, 0.8626, 0.0164, 0.8787, 0.8780, 0.0591, 0.8541, 0.8692,
        0.8905, 0.8559, 0.8723, 0.8523, 0.9583, 0.9486, 0.9922],
       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, 0.2274, 0.2043, 0.2064, 0.2125, 0.2022, 0.2049,
        0.2068, 0.2083, 0.2094, 0.2103, 0.2109, 0.2114, 0.2198])
finalReturns:  tensor([0.0397, 0.0478, 0.0354])
----------------------------------------
iter  1  stage  21  ep  29999   adversary:  const95-1.0,
  actions:  tensor([ 9,  9,  9, 12,  9,  9,  9,  9,  9, 12,  9,  9,  9,  9,  9,
9,  9,  9,
         9, 12,  9,  9,  9,  9,  0])
loss=  tensor(0.0137, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.7975,
0.7975, 0.7975, 0.7975, 0.7975, 0.7975, 0.7975, 0.7975, 0.7975,
        0.7975, 0.7975, 0.7975, 0.7975, 0.7975, 0.7975, 0.7975, 0.7975, 0.7975,
        0.7975, 0.7975, 0.7975, 0.7975, 0.5704, 0.3657, 0.1770]) return=
64479.669562396586
probs of actions:  tensor([0.9159, 0.9412, 0.9358, 0.0256, 0.9201, 0.9246,
0.9016, 0.9215, 0.9276,
        0.0340, 0.9117, 0.9078, 0.8983, 0.9093, 0.9158, 0.9064, 0.9014, 0.9157,
        0.9270, 0.0320, 0.9053, 0.9155, 0.9631, 0.9623, 0.9863],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3143, 0.2999, 0.2768, 0.2600,
0.2478, 0.2388,
        0.2259, 0.2346, 0.2290, 0.2249, 0.2219, 0.2196, 0.2179, 0.2166, 0.2156,
        0.2149, 0.2081, 0.2211, 0.2190, 0.2175, 0.2163, 0.2235])
finalReturns:  tensor([0.0788, 0.0869, 0.0741, 0.0465])
----------------------------------------
iter  1  stage  21  ep  59999   adversary:  const95-1.0,
  actions:  tensor([ 9,  9,  9,  9,  9,  9,  9,  9, 12,  9,  9, 12,  9,  9,  9,
9,  9,  6,
         9,  0,  9,  9,  9,  6,  0])
loss=  tensor(0.2748, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.7363,
0.7363, 0.7363, 0.7363, 0.7363, 0.7363, 0.7363, 0.7363, 0.7363,
        0.7363, 0.7363, 0.7363, 0.7363, 0.7363, 0.7363, 0.7363, 0.7363, 0.7363,
        0.7363, 0.7363, 0.7363, 0.7363, 0.5326, 0.3447, 0.1682]) return=
63525.18956622784
probs of actions:  tensor([0.9149, 0.9400, 0.9339, 0.9228, 0.9178, 0.9228,
0.9010, 0.9185, 0.0332,
        0.8975, 0.9112, 0.0388, 0.8970, 0.9078, 0.9157, 0.9048, 0.9007, 0.0332,
        0.9261, 0.0118, 0.9043, 0.9107, 0.9565, 0.0225, 0.9908],
       grad_fn=<ExpBackward0>)
```

```
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2302,
        0.2378, 0.2314, 0.2204, 0.2305, 0.2260, 0.2226, 0.2202, 0.2183, 0.2214,
        0.2089, 0.2179, 0.1900, 0.1956, 0.1998, 0.2075, 0.2067])
finalReturns:  tensor([0.0734, 0.0815, 0.0695, 0.0385])
----------------------------------------
iter  1  stage  21  ep  89999   adversary:  const95-1.0,
  actions:  tensor([ 9,  9,  9,  9, 12,  9,  9,  6,  9,  9,  9,  9,  9,  9,  9,
9,  9,  9,
         9,  9,  9,  9,  9,  9,  0])
loss=  tensor(0.0129, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.7825,
0.7825, 0.7825, 0.7825, 0.7825, 0.7825, 0.7825, 0.7825, 0.7825, 0.7825,
        0.7825, 0.7825, 0.7825, 0.7825, 0.7825, 0.7825, 0.7825, 0.7825, 0.7825,
        0.7825, 0.7825, 0.7825, 0.7825, 0.5611, 0.3605, 0.1749]) return=
63854.30325636411
probs of actions:  tensor([0.9253, 0.9474, 0.9409, 0.9324, 0.0420, 0.9313,
0.9130, 0.0223, 0.9336,
        0.9088, 0.9203, 0.9165, 0.9102, 0.9164, 0.9261, 0.9145, 0.9124, 0.9255,
        0.9341, 0.9146, 0.9132, 0.9181, 0.9589, 0.9694, 0.9937],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2854, 0.2788, 0.2615,
0.2534, 0.2322,
        0.2273, 0.2236, 0.2209, 0.2188, 0.2173, 0.2162, 0.2153, 0.2147, 0.2142,
        0.2139, 0.2136, 0.2134, 0.2133, 0.2131, 0.2131, 0.2211])
finalReturns:  tensor([0.0781, 0.0862, 0.0736, 0.0462])
----------------------------------------
iter  1  stage  21  ep  119999   adversary:  const95-1.0,
  actions:  tensor([ 9,  9,  9,  9,  9,  9,  9,  9,  9,  9,  9,  9,  9,  9,  9,
9,  9,  9,
         9, 12,  9,  9,  9,  9,  0])
loss=  tensor(0.0193, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.7966,
0.7966, 0.7966, 0.7966, 0.7966, 0.7966, 0.7966, 0.7966, 0.7966, 0.7966,
        0.7966, 0.7966, 0.7966, 0.7966, 0.7966, 0.7966, 0.7966, 0.7966, 0.7966,
        0.7966, 0.7966, 0.7966, 0.7966, 0.5698, 0.3654, 0.1769]) return=
64011.61156289775
probs of actions:  tensor([0.9028, 0.9281, 0.9189, 0.9113, 0.9003, 0.9093,
0.8888, 0.9008, 0.9104,
        0.8831, 0.8980, 0.8920, 0.8875, 0.8917, 0.9071, 0.8885, 0.8897, 0.9041,
        0.9133, 0.0715, 0.8885, 0.8815, 0.9319, 0.9598, 0.9956],
       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, 0.2075, 0.2206, 0.2187, 0.2172, 0.2161, 0.2234])
finalReturns:  tensor([0.0788, 0.0869, 0.0741, 0.0465])
----------------------------------------
iter  1  stage  21  ep  149999   adversary:  const95-1.0,
  actions:  tensor([ 9,  9,  9,  9,  9,  9,  9,  9, 12,  9,  9, 15, 12,  9,  9,
```

```
12,  9, 12,
         9,  9,  9,  9,  9,  9,  0])
loss=  tensor(0.0421, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.8001,
0.8001, 0.8001, 0.8001, 0.8001, 0.8001, 0.8001, 0.8001, 0.8001,
        0.8001, 0.8001, 0.8001, 0.8001, 0.8001, 0.8001, 0.8001, 0.8001, 0.8001,
        0.8001, 0.8001, 0.8001, 0.8001, 0.5720, 0.3666, 0.1774]) return=
65108.245685588154
probs of actions:  tensor([0.8282, 0.8623, 0.8464, 0.8418, 0.8199, 0.8360,
0.8128, 0.8229, 0.1314,
        0.8013, 0.8239, 0.0089, 0.1378, 0.8145, 0.8448, 0.1504, 0.8174, 0.1274,
        0.8468, 0.8121, 0.8161, 0.7587, 0.8521, 0.9195, 0.9976],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4193, 0.3613, 0.3206, 0.2917, 0.2708, 0.2557,
0.2446, 0.2302,
        0.2378, 0.2314, 0.2123, 0.2315, 0.2389, 0.2322, 0.2210, 0.2309, 0.2200,
        0.2302, 0.2258, 0.2225, 0.2200, 0.2182, 0.2169, 0.2239])
finalReturns:  tensor([0.0789, 0.0870, 0.0742, 0.0465])
----------------------------------------
iter  1  stage  21  ep  179999   adversary:  const95-1.0,
  actions:  tensor([12, 12, 12,  9,  9,  9,  9,  9,  9, 12,  9, 12,  9,  9, 12,
12, 12,  9,
        15, 12,  9,  9, 12,  9,  0])
loss=  tensor(0.1889, grad_fn=<NegBackward0>)   ,  base rewards= tensor([0.8339,
0.8339, 0.8339, 0.8339, 0.8339, 0.8339, 0.8339, 0.8339, 0.8339,
        0.8339, 0.8339, 0.8339, 0.8339, 0.8339, 0.8339, 0.8339, 0.8339, 0.8339,
        0.8339, 0.8339, 0.8339, 0.8339, 0.5927, 0.3780, 0.1822]) return=
66273.57654399677
probs of actions:  tensor([0.3208, 0.2918, 0.3097, 0.6656, 0.6256, 0.6527,
0.6345, 0.6362, 0.6539,
        0.3379, 0.6420, 0.3300, 0.6465, 0.6312, 0.2827, 0.3408, 0.3112, 0.6632,
        0.0122, 0.3303, 0.6451, 0.4837, 0.3308, 0.7926, 0.9987],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3408, 0.3061, 0.2812, 0.2633,
0.2502, 0.2405,
        0.2272, 0.2356, 0.2235, 0.2328, 0.2277, 0.2176, 0.2221, 0.2254, 0.2342,
        0.2144, 0.2331, 0.2401, 0.2331, 0.2216, 0.2314, 0.2348])
finalReturns:  tensor([0.0870, 0.0951, 0.0881, 0.0525])
----------------------------------------
iter  1  stage  20  ep  29999   adversary:  const95-1.0,
  actions:  tensor([ 9, 12, 12, 12,  9, 12, 12, 15, 12,  9,  9, 12,  9, 12, 12,
9, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.2135, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.0212,
1.0212, 1.0212, 1.0212, 1.0212, 1.0212, 1.0212, 1.0212, 1.0212,
        1.0212, 1.0212, 1.0212, 1.0212, 1.0212, 1.0212, 1.0212, 1.0212, 1.0212,
        1.0212, 1.0212, 1.0212, 0.7747, 0.5564, 0.3579, 0.1738]) return=
67469.79672706983
probs of actions:  tensor([0.2699, 0.6926, 0.7001, 0.6739, 0.2550, 0.6957,
```

```
0.6790, 0.0173, 0.6889,
        0.2628, 0.2843, 0.6976, 0.2999, 0.6886, 0.6446, 0.2603, 0.6645, 0.6650,
        0.6531, 0.6966, 0.6835, 0.8314, 0.7349, 0.5680, 0.9983],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4130, 0.3642, 0.3296, 0.3110, 0.2784, 0.2674,
0.2512, 0.2611,
        0.2609, 0.2484, 0.2330, 0.2399, 0.2267, 0.2289, 0.2369, 0.2244, 0.2272,
        0.2293, 0.2309, 0.2320, 0.2329, 0.2336, 0.2341, 0.2489])
finalReturns:  tensor([0.1603, 0.1747, 0.1602, 0.1251, 0.0751])
----------------------------------------
iter  1  stage  20  ep  59999   adversary:  const95-1.0,
  actions:  tensor([12, 12, 12, 12, 12,  9, 12, 12, 12, 12,  9, 12, 12, 12,  9,
12, 12,  9,
        12, 12, 12, 12,  9, 12,  0])
loss=  tensor(0.3982, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.0143,
1.0143, 1.0143, 1.0143, 1.0143, 1.0143, 1.0143, 1.0143, 1.0143,
        1.0143, 1.0143, 1.0143, 1.0143, 1.0143, 1.0143, 1.0143, 1.0143, 1.0143,
        1.0143, 1.0143, 1.0143, 0.7703, 0.5536, 0.3564, 0.1731]) return=
67733.46532960936
probs of actions:  tensor([0.8418, 0.8431, 0.8455, 0.8234, 0.8511, 0.1417,
0.8230, 0.8359, 0.8338,
        0.8354, 0.1529, 0.8379, 0.8105, 0.8279, 0.1763, 0.8419, 0.8088, 0.1605,
        0.8070, 0.8347, 0.8422, 0.9195, 0.1189, 0.7536, 0.9979],
       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, 0.2517, 0.2354, 0.2355, 0.2355, 0.2418, 0.2281, 0.2300, 0.2377,
        0.2250, 0.2276, 0.2296, 0.2311, 0.2385, 0.2257, 0.2425])
finalReturns:  tensor([0.1531, 0.1675, 0.1531, 0.1118, 0.0694])
----------------------------------------
iter  1  stage  20  ep  89999   adversary:  const95-1.0,
  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(0.0534, 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([0.9099, 0.9138, 0.9140, 0.8978, 0.9148, 0.9090,
0.8954, 0.9049, 0.9047,
        0.9035, 0.8959, 0.9057, 0.8875, 0.8993, 0.8838, 0.9082, 0.8855, 0.8936,
        0.8859, 0.9043, 0.9167, 0.9566, 0.9230, 0.8634, 0.9973],
       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  1  stage  20  ep  119999   adversary:  const95-1.0,
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 15,  9, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0448, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.0323,
1.0323, 1.0323, 1.0323, 1.0323, 1.0323, 1.0323, 1.0323, 1.0323, 1.0323,
        1.0323, 1.0323, 1.0323, 1.0323, 1.0323, 1.0323, 1.0323, 1.0323, 1.0323,
        1.0323, 1.0323, 1.0323, 0.7819, 0.5608, 0.3604, 0.1748]) return=
68680.34770364207
probs of actions:  tensor([0.9237, 0.9274, 0.9270, 0.9128, 0.9279, 0.9232,
0.9108, 0.9185, 0.9195,
        0.9184, 0.9107, 0.9201, 0.0090, 0.0734, 0.9014, 0.9213, 0.9018, 0.9092,
        0.9019, 0.9186, 0.9349, 0.9651, 0.9375, 0.8722, 0.9980],
       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, 0.2344, 0.2547, 0.2376, 0.2371, 0.2367, 0.2364,
        0.2362, 0.2361, 0.2359, 0.2359, 0.2358, 0.2357, 0.2501])
finalReturns:  tensor([0.1612, 0.1756, 0.1609, 0.1255, 0.0753])
----------------------------------------
iter  1  stage  20  ep  149999   adversary:  const95-1.0,
  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(0.0309, 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, 1.0332, 0.7825, 0.5612, 0.3606, 0.1749]) return=
68694.09895647192
probs of actions:  tensor([0.9452, 0.9495, 0.9486, 0.9376, 0.9484, 0.9455,
0.9354, 0.9409, 0.9425,
        0.9408, 0.9347, 0.9422, 0.9302, 0.9372, 0.9278, 0.9434, 0.9283, 0.9348,
        0.9291, 0.9408, 0.9553, 0.9748, 0.9568, 0.9097, 0.9988],
       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  1  stage  20  ep  179999   adversary:  const95-1.0,
  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(0.0189, 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, 0.7825, 0.5612, 0.3606, 0.1749]) return=
68694.09895647192
probs of actions:  tensor([0.9643, 0.9682, 0.9674, 0.9595, 0.9663, 0.9648,
0.9575, 0.9611, 0.9624,
        0.9606, 0.9569, 0.9619, 0.9534, 0.9585, 0.9525, 0.9628, 0.9520, 0.9574,
        0.9532, 0.9609, 0.9724, 0.9833, 0.9721, 0.9466, 0.9995],
       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  1  stage  20  ep  209999   adversary:  const95-1.0,
  actions:  tensor([ 9, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
15, 12, 12,
         9, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0135, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.0239,
1.0239, 1.0239, 1.0239, 1.0239, 1.0239, 1.0239, 1.0239, 1.0239,
        1.0239, 1.0239, 1.0239, 1.0239, 1.0239, 1.0239, 1.0239, 1.0239, 1.0239,
        1.0239, 1.0239, 1.0239, 0.7765, 0.5575, 0.3585, 0.1740]) return=
68418.15387194039
probs of actions:  tensor([0.0213, 0.9771, 0.9759, 0.9698, 0.9749, 0.9739,
0.9680, 0.9708, 0.9719,
        0.9702, 0.9677, 0.9714, 0.9648, 0.9688, 0.9647, 0.0047, 0.9634, 0.9685,
        0.0284, 0.9710, 0.9806, 0.9873, 0.9793, 0.9627, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.5031, 0.4130, 0.3642, 0.3296, 0.3047, 0.2866, 0.2734,
0.2637, 0.2565,
        0.2512, 0.2473, 0.2443, 0.2421, 0.2405, 0.2393, 0.2302, 0.2452, 0.2428,
        0.2473, 0.2321, 0.2330, 0.2337, 0.2341, 0.2345, 0.2492])
finalReturns:  tensor([0.1605, 0.1749, 0.1604, 0.1252, 0.0752])
----------------------------------------
iter  1  stage  20  ep  239999   adversary:  const95-1.0,
  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(0.0093, 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([0.9810, 0.9838, 0.9830, 0.9785, 0.9821, 0.9814,
0.9769, 0.9789, 0.9797,
        0.9782, 0.9766, 0.9792, 0.9742, 0.9773, 0.9744, 0.9799, 0.9734, 0.9771,
        0.9746, 0.9788, 0.9866, 0.9908, 0.9859, 0.9750, 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  1  stage  19  ep  29999   adversary:  const95-1.0,
  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(0.0147, 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, 0.9500, 0.7285, 0.5277, 0.3420, 0.1670]) return=
68694.09895647192
probs of actions:  tensor([0.9829, 0.9857, 0.9848, 0.9809, 0.9839, 0.9834,
0.9793, 0.9810, 0.9817,
        0.9800, 0.9791, 0.9810, 0.9765, 0.9796, 0.9771, 0.9821, 0.9761, 0.9795,
        0.9773, 0.9816, 0.9874, 0.9913, 0.9874, 0.9803, 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  1  stage  19  ep  59999   adversary:  const95-1.0,
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12,  9, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0115, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.1881,
1.1881, 1.1881, 1.1881, 1.1881, 1.1881, 1.1881, 1.1881, 1.1881,
        1.1881, 1.1881, 1.1881, 1.1881, 1.1881, 1.1881, 1.1881, 1.1881, 1.1881,
        1.1881, 1.1881, 0.9414, 0.7229, 0.5243, 0.3400, 0.1662]) return=
68487.85710629346
probs of actions:  tensor([0.9863, 0.9887, 0.9880, 0.9849, 0.9871, 0.9869,
0.9833, 0.9847, 0.9852,
        0.9837, 0.9833, 0.9847, 0.9810, 0.9835, 0.9818, 0.9857, 0.0142, 0.9837,
        0.9816, 0.9863, 0.9894, 0.9925, 0.9901, 0.9853, 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, 0.2441, 0.2298,
        0.2312, 0.2323, 0.2331, 0.2337, 0.2342, 0.2346, 0.2492])
finalReturns:  tensor([0.2290, 0.2434, 0.2289, 0.1937, 0.1437, 0.0830])
----------------------------------------
iter  1  stage  19  ep  89999   adversary:  const95-1.0,
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,  9, 12, 12,
```

```
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0109, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.1969,
1.1969, 1.1969, 1.1969, 1.1969, 1.1969, 1.1969, 1.1969, 1.1969,
        1.1969, 1.1969, 1.1969, 1.1969, 1.1969, 1.1969, 1.1969, 1.1969, 1.1969,
        1.1969, 1.1969, 0.9473, 0.7267, 0.5266, 0.3414, 0.1668]) return=
68466.13679373146
probs of actions:  tensor([0.9871, 0.9894, 0.9886, 0.9857, 0.9878, 0.9876,
0.9842, 0.9854, 0.9859,
        0.9844, 0.9841, 0.9854, 0.0124, 0.9845, 0.9828, 0.9864, 0.9816, 0.9844,
        0.9827, 0.9874, 0.9904, 0.9920, 0.9903, 0.9871, 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, 0.2488, 0.2332, 0.2338, 0.2343, 0.2346, 0.2348,
        0.2350, 0.2352, 0.2353, 0.2354, 0.2354, 0.2355, 0.2499])
finalReturns:  tensor([0.2297, 0.2441, 0.2295, 0.1941, 0.1440, 0.0831])
----------------------------------------
iter  1  stage  19  ep  119999   adversary:  const95-1.0,
  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(0.0119, 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, 0.9500, 0.7285, 0.5277, 0.3420, 0.1670]) return=
68694.09895647192
probs of actions:  tensor([0.9861, 0.9886, 0.9877, 0.9847, 0.9868, 0.9867,
0.9830, 0.9844, 0.9848,
        0.9831, 0.9829, 0.9842, 0.9807, 0.9830, 0.9819, 0.9855, 0.9805, 0.9834,
        0.9816, 0.9860, 0.9886, 0.9912, 0.9900, 0.9871, 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  1  stage  19  ep  149999   adversary:  const95-1.0,
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12,  9, 12, 12, 12,  0])
loss=  tensor(1.3077, 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, 0.9500, 0.7285, 0.5277, 0.3420, 0.1670]) return=
68553.0200585037
probs of actions:  tensor([0.9907, 0.9926, 0.9919, 0.9898, 0.9912, 0.9911,
```

```
0.9885, 0.9896, 0.9898,
        0.9885, 0.9884, 0.9893, 0.9868, 0.9885, 0.9877, 0.9902, 0.9867, 0.9888,
        0.9876, 0.9906, 0.0035, 0.9945, 0.9941, 0.9918, 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, 0.2426, 0.2287, 0.2304, 0.2317, 0.2471])
finalReturns:  tensor([0.2160, 0.2304, 0.2093, 0.1814, 0.1368, 0.0800])
----------------------------------------
iter  1  stage  19  ep  179999   adversary:  const95-1.0,
  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(0.0061, 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, 0.9500, 0.7285, 0.5277, 0.3420, 0.1670]) return=
68694.09895647192
probs of actions:  tensor([0.9924, 0.9940, 0.9934, 0.9917, 0.9928, 0.9928,
0.9905, 0.9914, 0.9916,
        0.9904, 0.9904, 0.9912, 0.9891, 0.9904, 0.9901, 0.9920, 0.9890, 0.9909,
        0.9898, 0.9925, 0.9938, 0.9950, 0.9955, 0.9941, 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  1  stage  19  ep  209999   adversary:  const95-1.0,
  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(0.0064, 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, 0.9500, 0.7285, 0.5277, 0.3420, 0.1670]) return=
68694.09895647192
probs of actions:  tensor([0.9923, 0.9939, 0.9933, 0.9915, 0.9926, 0.9926,
0.9902, 0.9912, 0.9913,
        0.9900, 0.9902, 0.9909, 0.9888, 0.9902, 0.9899, 0.9918, 0.9888, 0.9907,
        0.9896, 0.9921, 0.9934, 0.9947, 0.9953, 0.9946, 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  1  stage  19  ep  239999   adversary:  const95-1.0,
  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(0.0052, 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, 0.9500, 0.7285, 0.5277, 0.3420, 0.1670]) return=
68694.09895647192
probs of actions:  tensor([0.9937, 0.9951, 0.9945, 0.9930, 0.9939, 0.9940,
0.9920, 0.9928, 0.9928,
        0.9917, 0.9919, 0.9925, 0.9907, 0.9918, 0.9918, 0.9932, 0.9907, 0.9924,
        0.9915, 0.9935, 0.9943, 0.9953, 0.9964, 0.9963, 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  1  stage  19  ep  269999   adversary:  const95-1.0,
  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(0.0042, 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, 0.9500, 0.7285, 0.5277, 0.3420, 0.1670]) return=
68694.09895647192
probs of actions:  tensor([0.9949, 0.9961, 0.9956, 0.9943, 0.9950, 0.9951,
0.9934, 0.9942, 0.9941,
        0.9931, 0.9934, 0.9938, 0.9924, 0.9933, 0.9934, 0.9945, 0.9925, 0.9938,
        0.9930, 0.9945, 0.9955, 0.9960, 0.9971, 0.9974, 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  1  stage  19  ep  299999   adversary:  const95-1.0,
  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([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, 0.9965, 0.9961, 0.9950, 0.9956, 0.9957,
0.9942, 0.9949, 0.9947,
        0.9938, 0.9941, 0.9945, 0.9932, 0.9940, 0.9942, 0.9951, 0.9933, 0.9945,
        0.9938, 0.9955, 0.9958, 0.9966, 0.9976, 0.9978, 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  1  stage  18  ep  29999   adversary:  const95-1.0,
  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(0.0060, 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([0.9956, 0.9966, 0.9962, 0.9951, 0.9957, 0.9958,
0.9943, 0.9950, 0.9948,
        0.9939, 0.9942, 0.9946, 0.9934, 0.9942, 0.9944, 0.9952, 0.9935, 0.9947,
        0.9950, 0.9955, 0.9961, 0.9962, 0.9977, 0.9982, 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  1  stage  18  ep  59999   adversary:  const95-1.0,
  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(0.0030, 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, 0.9982, 0.9980, 0.9974, 0.9977, 0.9978,
0.9969, 0.9973, 0.9972,
        0.9966, 0.9968, 0.9971, 0.9963, 0.9968, 0.9969, 0.9974, 0.9963, 0.9971,
        0.9975, 0.9977, 0.9981, 0.9982, 0.9989, 0.9991, 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  1  stage  18  ep  89999   adversary:  const95-1.0,
  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([1.3631,
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,
        1.3631, 1.1119, 0.8901, 0.6892, 0.5034, 0.3283, 0.1613]) return=
68694.09895647192
probs of actions:  tensor([0.9973, 0.9980, 0.9977, 0.9970, 0.9973, 0.9975,
0.9965, 0.9969, 0.9968,
        0.9961, 0.9964, 0.9967, 0.9959, 0.9964, 0.9966, 0.9971, 0.9959, 0.9968,
        0.9972, 0.9974, 0.9977, 0.9979, 0.9989, 0.9991, 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  1  stage  18  ep  119999   adversary:  const95-1.0,
  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(0.0022, 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,
        1.3631, 1.1119, 0.8901, 0.6892, 0.5034, 0.3283, 0.1613]) return=
68694.09895647192
probs of actions:  tensor([0.9981, 0.9986, 0.9984, 0.9979, 0.9982, 0.9983,
0.9975, 0.9978, 0.9977,
        0.9973, 0.9975, 0.9977, 0.9971, 0.9974, 0.9976, 0.9979, 0.9971, 0.9977,
        0.9980, 0.9986, 0.9984, 0.9985, 0.9993, 0.9995, 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  1  stage  18  ep  149999   adversary:  const95-1.0,
  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(0.0017, 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([0.9984, 0.9989, 0.9987, 0.9983, 0.9984, 0.9986,
0.9979, 0.9982, 0.9981,
        0.9977, 0.9979, 0.9980, 0.9975, 0.9978, 0.9980, 0.9983, 0.9975, 0.9981,
        0.9985, 0.9989, 0.9987, 0.9988, 0.9994, 0.9997, 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  1  stage  18  ep  179999   adversary:  const95-1.0,
  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(0.0014, 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([0.9987, 0.9991, 0.9989, 0.9985, 0.9987, 0.9988,
0.9982, 0.9985, 0.9984,
        0.9980, 0.9982, 0.9983, 0.9978, 0.9982, 0.9983, 0.9985, 0.9979, 0.9984,
        0.9988, 0.9992, 0.9989, 0.9990, 0.9995, 0.9998, 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  1  stage  18  ep  190012   adversary:  const95-1.0,
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 15, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0011, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.3810,
1.3810, 1.3810, 1.3810, 1.3810, 1.3810, 1.3810, 1.3810, 1.3810,
        1.3810, 1.3810, 1.3810, 1.3810, 1.3810, 1.3810, 1.3810, 1.3810, 1.3810,
        1.3810, 1.1241, 0.8984, 0.6947, 0.5068, 0.3302, 0.1621]) return=
68884.88646283273
probs of actions:  tensor([0.9989, 0.9992, 0.9990, 0.9987, 0.9988, 0.9989,
```

```
0.9984, 0.9987, 0.9986,
        0.9982, 0.9984, 0.9985, 0.9981, 0.9984, 0.9985, 0.9987, 0.0012, 0.9986,
        0.9990, 0.9993, 0.9991, 0.9992, 0.9996, 0.9998, 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, 0.2297, 0.2448,
        0.2425, 0.2408, 0.2395, 0.2385, 0.2378, 0.2372, 0.2512])
finalReturns:  tensor([0.3064, 0.3208, 0.3058, 0.2700, 0.2195, 0.1582, 0.0892])
----------------------------------------
iter  1  stage  17  ep  1044   adversary:  const95-1.0,
  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(0.0017, 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([0.9989, 0.9992, 0.9991, 0.9987, 0.9989, 0.9990,
0.9985, 0.9987, 0.9986,
        0.9983, 0.9984, 0.9986, 0.9981, 0.9984, 0.9986, 0.9988, 0.9982, 0.9990,
        0.9990, 0.9993, 0.9991, 0.9992, 0.9996, 0.9998, 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  1  stage  16  ep  29999   adversary:  const95-1.0,
  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(0.0017, 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, 1.4248,
        1.2024, 1.0011, 0.8148, 0.6396, 0.4723, 0.3109, 0.1538]) return=
68694.09895647192
probs of actions:  tensor([0.9992, 0.9994, 0.9993, 0.9991, 0.9992, 0.9993,
0.9989, 0.9991, 0.9990,
        0.9987, 0.9989, 0.9990, 0.9986, 0.9989, 0.9990, 0.9991, 0.9989, 0.9994,
        0.9994, 0.9996, 0.9994, 0.9996, 0.9997, 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  1  stage  16  ep  35741   adversary:  const95-1.0,
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 15, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0015, grad_fn=<NegBackward0>)   ,  base rewards= tensor([1.6877,
1.6877, 1.6877, 1.6877, 1.6877, 1.6877, 1.6877, 1.6877, 1.6877,
        1.6877, 1.6877, 1.6877, 1.6877, 1.6877, 1.6877, 1.6877, 1.6877, 1.4323,
        1.2076, 1.0047, 0.8173, 0.6412, 0.4733, 0.3115, 0.1541]) return=
68906.62996763481
probs of actions:  tensor([9.9926e-01, 9.9949e-01, 9.9938e-01, 9.9916e-01,
9.9924e-01, 9.9932e-01,
        9.9897e-01, 9.9913e-01, 9.9906e-01, 9.9883e-01, 9.9895e-01, 9.9905e-01,
        8.9071e-04, 9.9894e-01, 9.9906e-01, 9.9918e-01, 9.9901e-01, 9.9949e-01,
        9.9950e-01, 9.9960e-01, 9.9945e-01, 9.9967e-01, 9.9976e-01, 9.9999e-01,
        1.0000e+00], 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, 0.2344, 0.2484, 0.2451, 0.2427, 0.2409, 0.2396,
        0.2386, 0.2378, 0.2373, 0.2369, 0.2365, 0.2363, 0.2505])
finalReturns:  tensor([0.4669, 0.4813, 0.4663, 0.4307, 0.3802, 0.3190, 0.2501,
0.1754, 0.0964])
---------------------------------------
iter  1  stage  15  ep  0   adversary:  const95-1.0,
  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(0.0021, 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, 1.5780, 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, 0.9995, 0.9994, 0.9992, 0.9992, 0.9993,
0.9990, 0.9991, 0.9991,
        0.9988, 0.9989, 0.9990, 0.9987, 0.9990, 0.9990, 0.9992, 0.9990, 0.9995,
        0.9995, 0.9996, 0.9995, 0.9997, 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  1  stage  14  ep  0   adversary:  const95-1.0,
  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(0.0030, 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, 0.9994, 0.9992, 0.9992, 0.9993,
0.9990, 0.9991, 0.9991,
        0.9988, 0.9989, 0.9990, 0.9987, 0.9990, 0.9990, 0.9992, 0.9990, 0.9995,
        0.9995, 0.9996, 0.9995, 0.9997, 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  1  stage  13  ep  60   adversary:  const95-1.0,
  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(0.0039, 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, 0.9994, 0.9992, 0.9992, 0.9993,
0.9990, 0.9991, 0.9991,
        0.9988, 0.9989, 0.9990, 0.9987, 0.9990, 0.9991, 0.9992, 0.9990, 0.9995,
        0.9995, 0.9996, 0.9994, 0.9997, 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  1  stage  12  ep  674   adversary:  const95-1.0,
  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([2.2903,
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, 0.9994, 0.9992, 0.9992, 0.9993,
0.9990, 0.9991, 0.9991,
        0.9988, 0.9989, 0.9991, 0.9990, 0.9992, 0.9993, 0.9993, 0.9990, 0.9995,
        0.9995, 0.9996, 0.9994, 0.9997, 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  1  stage  11  ep  0   adversary:  const95-1.0,
  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(0.0058, 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([0.9993, 0.9995, 0.9994, 0.9992, 0.9992, 0.9993,
0.9990, 0.9991, 0.9991,
        0.9988, 0.9989, 0.9991, 0.9990, 0.9992, 0.9993, 0.9993, 0.9990, 0.9995,
        0.9995, 0.9996, 0.9994, 0.9997, 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  1  stage  10  ep  58   adversary:  const95-1.0,
  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([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([0.9993, 0.9995, 0.9994, 0.9992, 0.9992, 0.9993,
0.9990, 0.9991, 0.9991,
        0.9988, 0.9990, 0.9991, 0.9990, 0.9992, 0.9993, 0.9993, 0.9990, 0.9995,
        0.9995, 0.9996, 0.9994, 0.9997, 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  1  stage  9  ep  794   adversary:  const95-1.0,
  actions:  tensor([12, 12, 12, 12, 12, 12, 12, 15, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12,
        12, 12, 12, 12, 12, 12,  0])
loss=  tensor(0.0075, grad_fn=<NegBackward0>)   ,  base rewards= tensor([2.7830,
2.7830, 2.7830, 2.7830, 2.7830, 2.7830, 2.7830, 2.7830, 2.7830,
        2.7830, 2.5107, 2.2742, 2.0629, 1.8694, 1.6889, 1.5178, 1.3536, 1.1945,
        1.0391, 0.8865, 0.7360, 0.5870, 0.4391, 0.2922, 0.1458]) return=
68919.5088490973
probs of actions:  tensor([9.9929e-01, 9.9951e-01, 9.9940e-01, 9.9919e-01,
9.9927e-01, 9.9935e-01,
        9.9900e-01, 6.2436e-04, 9.9908e-01, 9.9900e-01, 9.9922e-01, 9.9933e-01,
        9.9913e-01, 9.9936e-01, 9.9929e-01, 9.9932e-01, 9.9904e-01, 9.9950e-01,
        9.9951e-01, 9.9961e-01, 9.9946e-01, 9.9970e-01, 9.9976e-01, 1.0000e+00,
        1.0000e+00], grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3345, 0.3083, 0.2892, 0.2753,
0.2570, 0.2655,
        0.2578, 0.2522, 0.2480, 0.2449, 0.2425, 0.2408, 0.2395, 0.2385, 0.2378,
        0.2372, 0.2368, 0.2365, 0.2363, 0.2361, 0.2360, 0.2503])
finalReturns:  tensor([1.0883, 1.1027, 1.0870, 1.0504, 0.9989, 0.9369, 0.8672,
0.7919, 0.7126,
        0.6302, 0.5455, 0.4592, 0.3717, 0.2833, 0.1941, 0.1044])
----------------------------------------
iter  1  stage  8  ep  0   adversary:  const95-1.0,
  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([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([0.9993, 0.9995, 0.9994, 0.9992, 0.9993, 0.9993,
0.9990, 0.9992, 0.9991,
```

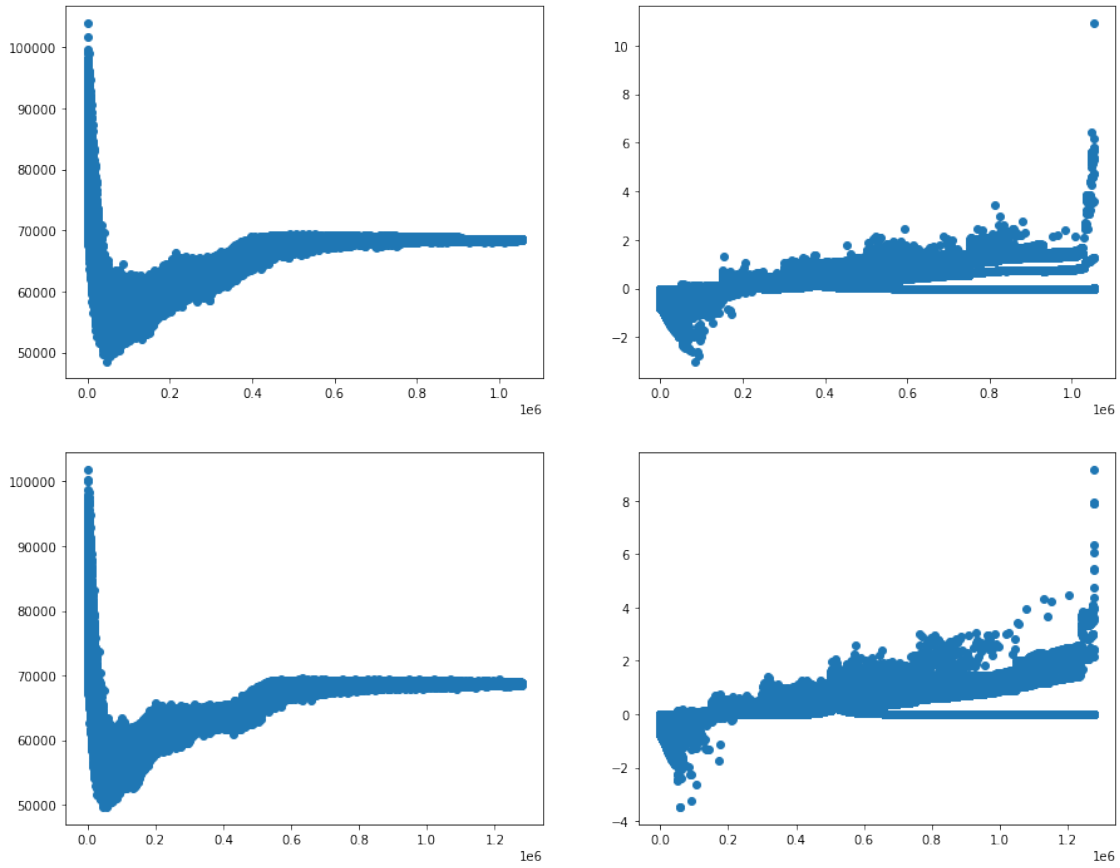
```
        0.9990, 0.9992, 0.9993, 0.9991, 0.9994, 0.9993, 0.9993, 0.9990, 0.9995,
        0.9995, 0.9996, 0.9995, 0.9997, 0.9998, 1.0000, 1.0000],
       grad_fn=<ExpBackward0>)
rewards:  tensor([0.4968, 0.4229, 0.3712, 0.3345, 0.3083, 0.2892, 0.2753,
0.2651, 0.2576,
        0.2520, 0.2479, 0.2448, 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  1  stage  7  ep  0   adversary:  const95-1.0,
  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(0.0103, 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([0.9993, 0.9995, 0.9994, 0.9992, 0.9993, 0.9993,
0.9990, 0.9992, 0.9991,
        0.9990, 0.9992, 0.9993, 0.9991, 0.9994, 0.9993, 0.9993, 0.9990, 0.9995,
        0.9995, 0.9996, 0.9995, 0.9997, 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  1  stage  6  ep  2   adversary:  const95-1.0,
  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(0.0120, 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([0.9993, 0.9995, 0.9994, 0.9992, 0.9993, 0.9993,
0.9990, 0.9992, 0.9991,
        0.9990, 0.9992, 0.9993, 0.9991, 0.9994, 0.9993, 0.9993, 0.9990, 0.9995,
        0.9995, 0.9996, 0.9995, 0.9997, 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  1  stage  5  ep  0   adversary:  const95-1.0,
  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([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([0.9993, 0.9995, 0.9994, 0.9992, 0.9993, 0.9993,
0.9990, 0.9992, 0.9991,
        0.9990, 0.9992, 0.9993, 0.9991, 0.9994, 0.9993, 0.9993, 0.9990, 0.9995,
        0.9995, 0.9996, 0.9995, 0.9997, 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  1  stage  4  ep  0   adversary:  const95-1.0,
  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(0.0148, 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([0.9993, 0.9995, 0.9994, 0.9992, 0.9993, 0.9993,
0.9990, 0.9992, 0.9991,
        0.9990, 0.9992, 0.9993, 0.9991, 0.9994, 0.9993, 0.9993, 0.9990, 0.9995,
        0.9995, 0.9996, 0.9995, 0.9997, 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  1  stage  3  ep  0   adversary:  const95-1.0,
  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(0.0165, 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([0.9993, 0.9995, 0.9994, 0.9992, 0.9993, 0.9993,
0.9990, 0.9992, 0.9991,
        0.9990, 0.9992, 0.9993, 0.9991, 0.9994, 0.9993, 0.9993, 0.9990, 0.9995,
        0.9995, 0.9996, 0.9995, 0.9997, 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  1  stage  2  ep  0   adversary:  const95-1.0,
  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(0.0178, 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([0.9993, 0.9995, 0.9994, 0.9992, 0.9993, 0.9993,
0.9990, 0.9992, 0.9991,
        0.9990, 0.9992, 0.9993, 0.9991, 0.9994, 0.9993, 0.9993, 0.9990, 0.9995,
        0.9995, 0.9996, 0.9995, 0.9997, 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  1  stage  1  ep  0   adversary:  const95-1.0,
  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(0.0191, 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([0.9993, 0.9995, 0.9994, 0.9992, 0.9993, 0.9993,
0.9990, 0.9992, 0.9991,
        0.9990, 0.9992, 0.9993, 0.9991, 0.9994, 0.9993, 0.9993, 0.9990, 0.9995,
        0.9995, 0.9996, 0.9995, 0.9997, 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  1  stage  0  ep  0   adversary:  const95-1.0,
  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(0.0208, 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([0.9993, 0.9995, 0.9994, 0.9992, 0.9993, 0.9993,
0.9990, 0.9992, 0.9991,
        0.9990, 0.9992, 0.9993, 0.9991, 0.9994, 0.9993, 0.9993, 0.9990, 0.9995,
        0.9995, 0.9996, 0.9995, 0.9997, 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])
policy saved!
1,3,[5e-06,1][1, 10000, 1, 1],1683196677 saved
```



[5]:

[6]:

[6]: []

[ ]: