```
Last login: Sun Mar 29 15:32:49 on ttys000
Run-Mac:~ mac$ cd ~/.ssh
Run-Mac:.ssh mac$ ssh -i "Runzhe.pem" ubuntu@ec2-34-200-226-196.compute-1.amazonaws.com
The authenticity of host 'ec2-34-200-226-196.compute-1.amazonaws.com (34.200.226.196)' can't be established.
ECDSA key fingerprint is SHA256:w+hExzKE0n8gWq/kqgcL/n3mfYBX1XYDeVMmprGcIbI.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'ec2-34-200-226-196.compute-1.amazonaws.com,34.200.226.196' (ECDSA) to the list of known hosts.
Welcome to Ubuntu 18.04.3 LTS (GNU/Linux 4.15.0-1060-aws x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

  System information as of Mon Mar 30 01:15:26 UTC 2020

  System load:  0.53               Processes:           219
  Usage of /:   55.4% of 15.45GB   Users logged in:     0
  Memory usage: 1%                 IP address for ens5: 172.31.15.241
  Swap usage:   0%

 * Kubernetes 1.18 GA is now available! See https://microk8s.io for docs or
   install it with:

     sudo snap install microk8s --channel=1.18 --classic

 * Multipass 1.1 adds proxy support for developers behind enterprise
   firewalls. Rapid prototyping for cloud operations just got easier.

     https://multipass.run/

 * Canonical Livepatch is available for installation.
   - Reduce system reboots and improve kernel security. Activate at:
     https://ubuntu.com/livepatch

53 packages can be updated.
0 updates are security updates.


Last login: Thu Mar  5 21:23:34 2020 from 107.13.161.147
ubuntu@ip-172-31-15-241:~$ export openblas_num_threads=1; export OMP_NUM_THREADS=1
ubuntu@ip-172-31-15-241:~$ python EC2.py
21:17, 03/29; num of cores:16

Basic setting:[sd_O, sd_D, sd_R, sd_u_O, w_O, w_A, lam] = [2, 2, None, 0.4, 1, 1, 0.0001]


--------------------------------------
[pattern_seed, T, sd_R] = [0, 672, 0]

max(u_O) =  27.327727595549877
O_threshold = 12
means of Order:

22.323 12.937 16.305 27.014 23.267

7.457 16.12 10.376 10.577 12.991

11.677 19.721 14.946 11.573 13.165

12.597 20.038 10.155 12.494 7.833

3.97 14.317 15.577 8.192 27.328

target policy:

1 1 1 1 1

0 1 0 0 1

0 1 1 0 1

1 1 0 1 0

0 1 1 0 1

number of reward locations:  16
O_threshold = 9
target policy:

1 1 1 1 1

0 1 1 1 1

1 1 1 1 1

1 1 1 1 0

0 1 1 0 1
```

```
number of reward locations:  21
O_threshold = 15
target policy:

1 0 1 1 1

0 1 0 0 0

0 1 0 0 0

0 1 0 0 0

0 0 1 0 1

number of reward locations:  9
1 2 3 1 2 3
--------------------------------------
O_threshold = 12
MC-based mean and std of average reward:[1.1718e+01 5.0000e-03]
Value of Behaviour policy:11.24
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.11, 0.11, 0.1]][[0.15, 0.14, 0.15]][[-11.72, -11.72, -11.72]][[0.1, -0.48]]
std:[[0.01, 0.0, 0.01]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.0, 0.01]]
MSE:[[0.11, 0.11, 0.1]][[0.15, 0.14, 0.15]][[11.72, 11.72, 11.72]][[0.1, 0.48]]
MSE(-DR):[[0.0, 0.0, -0.01]][[0.04, 0.03, 0.04]][[11.61, 11.61, 11.61]][[-0.01, 0.37]]
better than DR_NO_MARL
==============
O_threshold = 9
MC-based mean and std of average reward:[1.1523e+01 5.0000e-03]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.4, 0.39, 0.4]][[0.45, 0.44, 0.45]][[-11.52, -11.52, -11.52]][[0.39, -0.28]]
std:[[0.03, 0.03, 0.01]][[0.02, 0.02, 0.02]][[0.0, 0.0, 0.0]][[0.01, 0.01]]
MSE:[[0.4, 0.39, 0.4]][[0.45, 0.44, 0.45]][[11.52, 11.52, 11.52]][[0.39, 0.28]]
MSE(-DR):[[0.0, -0.01, 0.0]][[0.05, 0.04, 0.05]][[11.12, 11.12, 11.12]][[-0.01, -0.12]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
MC-based ATE = -0.2
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.28, 0.28, 0.29]][[0.3, 0.3, 0.3]][[0.2, 0.2, 0.2]][0.29]
std:[[0.04, 0.04, 0.02]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][0.01]
MSE:[[0.28, 0.28, 0.29]][[0.3, 0.3, 0.3]][[0.2, 0.2, 0.2]][0.29]
MSE(-DR):[[0.0, 0.0, 0.01]][[0.02, 0.02, 0.02]][[-0.08, -0.08, -0.08]][0.01]
***** BETTER THAN [IS, DR_NO_MARL] *****
==============
O_threshold = 15
MC-based mean and std of average reward:[1.1758e+01 4.0000e-03]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.26, -0.27, -0.22]][[-0.38, -0.38, -0.37]][[-11.76, -11.76, -11.76]][[-0.23, -0.52]]
std:[[0.01, 0.0, 0.0]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.0, 0.01]]
MSE:[[0.26, 0.27, 0.22]][[0.38, 0.38, 0.37]][[11.76, 11.76, 11.76]][[0.23, 0.52]]
MSE(-DR):[[0.0, 0.01, -0.04]][[0.12, 0.12, 0.11]][[11.5, 11.5, 11.5]][[-0.03, 0.26]]
better than DR_NO_MARL
MC-based ATE = 0.04
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.37, -0.37, -0.33]][[-0.53, -0.53, -0.52]][[-0.04, -0.04, -0.04]][-0.33]
std:[[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.0]
MSE:[[0.37, 0.37, 0.33]][[0.53, 0.53, 0.52]][[0.04, 0.04, 0.04]][0.33]
MSE(-DR):[[0.0, 0.0, -0.04]][[0.16, 0.16, 0.15]][[-0.33, -0.33, -0.33]][-0.04]
better than DR_NO_MARL
==============
time spent until now: 2.8 mins


--------------------------------------
[pattern_seed, T, sd_R] = [0, 672, 2]

max(u_O) =  27.327727595549877
O_threshold = 12
means of Order:

22.323 12.937 16.305 27.014 23.267

7.457 16.12 10.376 10.577 12.991

11.677 19.721 14.946 11.573 13.165

12.597 20.038 10.155 12.494 7.833

3.97 14.317 15.577 8.192 27.328

target policy:

1 1 1 1 1

0 1 0 0 1

0 1 1 0 1

1 1 0 1 0
```

```
0 1 1 0 1

number of reward locations:  16
O_threshold = 9
target policy:

1 1 1 1 1

0 1 1 1 1

1 1 1 1 1

1 1 1 1 0

0 1 1 0 1

number of reward locations:  21
O_threshold = 15
target policy:

1 0 1 1 1

0 1 0 0 0

0 1 0 0 0

0 1 0 0 0

0 0 1 0 1

number of reward locations:  9
1 2 3 1 2 3
----------------------------------------
O_threshold = 12
MC-based mean and std of average reward:[11.717  0.015]
Value of Behaviour policy:11.244
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.16, 0.15, 0.14]][[0.17, 0.16, 0.16]][[-11.72, -11.72, -11.72]][[0.13, -0.47]]
std:[[0.01, 0.0, 0.0]][[0.02, 0.02, 0.02]][[0.0, 0.0, 0.0]][[0.0, 0.01]]
MSE:[[0.16, 0.15, 0.14]][[0.17, 0.16, 0.16]][[11.72, 11.72, 11.72]][[0.13, 0.47]]
MSE(-DR):[[0.0, -0.01, -0.02]][[0.01, 0.0, 0.0]][[11.56, 11.56, 11.56]][[-0.03, 0.31]]
better than DR_NO_MARL
==============
O_threshold = 9
MC-based mean and std of average reward:[11.523  0.016]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.45, 0.45, 0.43]][[0.46, 0.45, 0.46]][[-11.52, -11.52, -11.52]][[0.42, -0.28]]
std:[[0.02, 0.02, 0.01]][[0.03, 0.02, 0.02]][[0.0, 0.0, 0.0]][[0.01, 0.01]]
MSE:[[0.45, 0.45, 0.43]][[0.46, 0.45, 0.46]][[11.52, 11.52, 11.52]][[0.42, 0.28]]
MSE(-DR):[[0.0, 0.0, -0.02]][[0.01, 0.0, 0.01]][[11.07, 11.07, 11.07]][[-0.03, -0.17]]
better than DR_NO_MARL
MC-based ATE = -0.19
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.3, 0.29, 0.29]][[0.29, 0.29, 0.3]][[0.19, 0.19, 0.19]][0.29]
std:[[0.01, 0.02, 0.01]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.01]
MSE:[[0.3, 0.29, 0.29]][[0.29, 0.29, 0.3]][[0.19, 0.19, 0.19]][0.29]
MSE(-DR):[[0.0, -0.01, -0.01]][[-0.01, -0.01, 0.0]][[-0.11, -0.11, -0.11]][-0.01]
==============
O_threshold = 15
MC-based mean and std of average reward:[11.758  0.015]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.18, -0.19, -0.18]][[-0.36, -0.36, -0.36]][[-11.76, -11.76, -11.76]][[-0.19, -0.51]]
std:[[0.02, 0.01, 0.01]][[0.03, 0.02, 0.02]][[0.0, 0.0, 0.0]][[0.0, 0.01]]
MSE:[[0.18, 0.19, 0.18]][[0.36, 0.36, 0.36]][[11.76, 11.76, 11.76]][[0.19, 0.51]]
MSE(-DR):[[0.0, 0.01, 0.0]][[0.18, 0.18, 0.18]][[11.58, 11.58, 11.58]][[0.01, 0.33]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
MC-based ATE = 0.04
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.33, -0.34, -0.31]][[-0.52, -0.52, -0.52]][[-0.04, -0.04, -0.04]][-0.32]
std:[[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.0]
MSE:[[0.33, 0.34, 0.31]][[0.52, 0.52, 0.52]][[0.04, 0.04, 0.04]][0.32]
MSE(-DR):[[0.0, 0.01, -0.02]][[0.19, 0.19, 0.19]][[-0.29, -0.29, -0.29]][-0.01]
better than DR_NO_MARL
==============
time spent until now: 5.2 mins


----------------------------------------
[pattern_seed, T, sd_R] = [1, 672, 0]

max(u_O) =  22.15193176791189
O_threshold = 12
means of Order:

21.11 8.63 8.924 7.177 15.583

4.39 22.152 8.13 12.524 9.977

19.783 4.835 9.689 9.453 17.349
```

```
7.1 10.289 7.759 11.211 13.917

7.098 17.425 15.81 13.477 15.805

target policy:

1 0 0 0 1

0 1 0 1 0

1 0 0 0 1

0 0 0 0 1

0 1 1 1 1

number of reward locations:  11
O_threshold = 9
target policy:

1 0 0 0 1

0 1 0 1 1

1 0 1 1 1

0 1 0 1 1

0 1 1 1 1

number of reward locations:  16
O_threshold = 15
target policy:

1 0 0 0 1

0 1 0 0 0

1 0 0 0 1

0 0 0 0 0

0 1 1 0 1

number of reward locations:  8
1 2 3 1 2 3
----------------------------------------
O_threshold = 12
MC-based mean and std of average reward:[9.295e+00 5.000e-03]
Value of Behaviour policy:8.886
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.08, -0.08, -0.06]][[-0.16, -0.16, -0.15]][[-9.3, -9.3, -9.3]][[-0.07, -0.41]]
std:[[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.0, 0.0]]
MSE:[[0.08, 0.08, 0.06]][[0.16, 0.16, 0.15]][[9.3, 9.3, 9.3]][[0.07, 0.41]]
MSE(-DR):[[0.0, 0.0, -0.02]][[0.08, 0.08, 0.07]][[9.22, 9.22, 9.22]][[-0.01, 0.33]]
better than DR_NO_MARL
==============
O_threshold = 9
MC-based mean and std of average reward:[9.2e+00 6.0e-03]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.19, 0.18, 0.2]][[0.16, 0.15, 0.16]][[-9.2, -9.2, -9.2]][[0.19, -0.31]]
std:[[0.01, 0.01, 0.02]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.02, 0.0]]
MSE:[[0.19, 0.18, 0.2]][[0.16, 0.15, 0.16]][[9.2, 9.2, 9.2]][[0.19, 0.31]]
MSE(-DR):[[0.0, -0.01, 0.01]][[-0.03, -0.04, -0.03]][[9.01, 9.01, 9.01]][[0.0, 0.12]]
MC-based ATE = -0.1
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.27, 0.27, 0.26]][[0.31, 0.31, 0.31]][[0.1, 0.1, 0.1]][0.26]
std:[[0.01, 0.01, 0.01]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][0.01]
MSE:[[0.27, 0.27, 0.26]][[0.31, 0.31, 0.31]][[0.1, 0.1, 0.1]][0.26]
MSE(-DR):[[0.0, 0.0, -0.01]][[0.04, 0.04, 0.04]][[-0.17, -0.17, -0.17]][-0.01]
better than DR_NO_MARL
==============
O_threshold = 15
MC-based mean and std of average reward:[9.261e+00 5.000e-03]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.23, -0.24, -0.18]][[-0.36, -0.36, -0.35]][[-9.26, -9.26, -9.26]][[-0.19, -0.38]]
std:[[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.0, 0.0]]
MSE:[[0.23, 0.24, 0.18]][[0.36, 0.36, 0.35]][[9.26, 9.26, 9.26]][[0.19, 0.38]]
MSE(-DR):[[0.0, 0.01, -0.05]][[0.13, 0.13, 0.12]][[9.03, 9.03, 9.03]][[-0.04, 0.15]]
better than DR_NO_MARL
MC-based ATE = -0.03
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.16, -0.16, -0.12]][[-0.2, -0.2, -0.2]][[0.03, 0.03, 0.03]][-0.13]
std:[[0.01, 0.0, 0.01]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.0]
MSE:[[0.16, 0.16, 0.12]][[0.2, 0.2, 0.2]][[0.03, 0.03, 0.03]][0.13]
MSE(-DR):[[0.0, 0.0, -0.04]][[0.04, 0.04, 0.04]][[-0.13, -0.13, -0.13]][-0.03]
better than DR_NO_MARL
==============
```

```
time spent until now: 7.7 mins


---------------------------------------
[pattern_seed, T, sd_R] = [1, 672, 2]

max(u_O) =  22.15193176791189
O_threshold = 12
means of Order:

21.11 8.63 8.924 7.177 15.583

4.39 22.152 8.13 12.524 9.977

19.783 4.835 9.689 9.453 17.349

7.1 10.289 7.759 11.211 13.917

7.098 17.425 15.81 13.477 15.805

target policy:

1 0 0 0 1

0 1 0 1 0

1 0 0 0 1

0 0 0 0 1

0 1 1 1 1

number of reward locations:  11
O_threshold = 9
target policy:

1 0 0 0 1

0 1 0 1 1

1 0 1 1 1

0 1 0 1 1

0 1 1 1 1

number of reward locations:  16
O_threshold = 15
target policy:

1 0 0 0 1

0 1 0 0 0

1 0 0 0 1

0 0 0 0 0

0 1 1 0 1

number of reward locations:  8
1 2 3 1 2 3
---------------------------------------
O_threshold = 12
MC-based mean and std of average reward:[9.294 0.016]
Value of Behaviour policy:8.889
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.07, -0.08, -0.04]][[-0.14, -0.14, -0.14]][[-9.29, -9.29, -9.29]][[-0.05, -0.41]]
std:[[0.08, 0.08, 0.05]][[0.02, 0.02, 0.02]][[0.0, 0.0, 0.0]][[0.05, 0.0]]
MSE:[[0.11, 0.11, 0.06]][[0.14, 0.14, 0.14]][[9.29, 9.29, 9.29]][[0.07, 0.41]]
MSE(-DR):[[0.0, 0.0, -0.05]][[0.03, 0.03, 0.03]][[9.18, 9.18, 9.18]][[-0.04, 0.3]]
better than DR_NO_MARL
==============
O_threshold = 9
MC-based mean and std of average reward:[9.2   0.016]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.17, 0.17, 0.21]][[0.17, 0.17, 0.17]][[-9.2, -9.2, -9.2]][[0.21, -0.31]]
std:[[0.02, 0.02, 0.02]][[0.03, 0.03, 0.03]][[0.0, 0.0, 0.0]][[0.02, 0.0]]
MSE:[[0.17, 0.17, 0.21]][[0.17, 0.17, 0.17]][[9.2, 9.2, 9.2]][[0.21, 0.31]]
MSE(-DR):[[0.0, 0.0, 0.04]][[0.0, 0.0, 0.0]][[9.03, 9.03, 9.03]][[0.04, 0.14]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
MC-based ATE = -0.09
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.25, 0.25, 0.25]][[0.31, 0.31, 0.31]][[0.09, 0.09, 0.09]][0.26]
std:[[0.1, 0.1, 0.07]][[0.01, 0.01, 0.0]][[0.0, 0.0, 0.0]][0.07]
MSE:[[0.27, 0.27, 0.26]][[0.31, 0.31, 0.31]][[0.09, 0.09, 0.09]][0.27]
MSE(-DR):[[0.0, 0.0, -0.01]][[0.04, 0.04, 0.04]][[-0.18, -0.18, -0.18]][0.0]
better than DR_NO_MARL
==============
```

```
O_threshold = 15
MC-based mean and std of average reward:[9.26  0.016]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.2, -0.21, -0.15]][[-0.34, -0.35, -0.34]][[-9.26, -9.26, -9.26]][[-0.15, -0.37]]
std:[[0.03, 0.03, 0.02]][[0.02, 0.02, 0.02]][[0.0, 0.0, 0.0]][[0.02, 0.0]]
MSE:[[0.2, 0.21, 0.15]][[0.34, 0.35, 0.34]][[9.26, 9.26, 9.26]][[0.15, 0.37]]
MSE(-DR):[[0.0, 0.01, -0.05]][[0.14, 0.15, 0.14]][[9.06, 9.06, 9.06]][[-0.05, 0.17]]
better than DR_NO_MARL
MC-based ATE = -0.03
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.13, -0.13, -0.1]][[-0.2, -0.21, -0.2]][[0.03, 0.03, 0.03]][-0.1]
std:[[0.05, 0.05, 0.03]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.03]
MSE:[[0.14, 0.14, 0.1]][[0.2, 0.21, 0.2]][[0.03, 0.03, 0.03]][0.1]
MSE(-DR):[[0.0, 0.0, -0.04]][[0.06, 0.07, 0.06]][[-0.11, -0.11, -0.11]][-0.04]
better than DR_NO_MARL
==============
time spent until now: 10.1 mins


---------------------------------------
[pattern_seed, T, sd_R] = [2, 672, 0]

max(u_O) =  27.57427313220561
O_threshold = 12
means of Order:

9.331 10.778 4.69 21.245 5.38

7.872 13.479 6.699 7.22 7.663

13.744 27.574 11.208 7.049 13.676

8.684 10.939 17.637 8.173 11.063

7.758 10.355 12.215 7.422 9.626

target policy:

0 0 0 1 0

0 1 0 0 0

1 1 0 0 1

0 0 1 0 0

0 0 1 0 0

number of reward locations:  7
O_threshold = 9
target policy:

1 1 0 1 0

0 1 0 0 0

1 1 1 0 1

0 1 1 0 1

0 1 1 0 1

number of reward locations:  14
O_threshold = 15
target policy:

0 0 0 1 0

0 0 0 0 0

0 1 0 0 0

0 0 1 0 0

0 0 0 0 0

number of reward locations:  3
1 2 3 1 2 3
---------------------------------------
O_threshold = 12
MC-based mean and std of average reward:[8.426e+00 4.000e-03]
Value of Behaviour policy:8.118
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.28, -0.29, -0.26]][[-0.36, -0.36, -0.35]][[-8.43, -8.43, -8.43]][[-0.26, -0.31]]
std:[[0.01, 0.01, 0.0]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.0, 0.0]]
MSE:[[0.28, 0.29, 0.26]][[0.36, 0.36, 0.35]][[8.43, 8.43, 8.43]][[0.26, 0.31]]
MSE(-DR):[[0.0, 0.01, -0.02]][[0.08, 0.08, 0.07]][[8.15, 8.15, 8.15]][[-0.02, 0.03]]
better than DR_NO_MARL
==============
```

```
O_threshold = 9
MC-based mean and std of average reward:[8.467e+00 4.000e-03]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.03, 0.02, 0.03]][[0.03, 0.03, 0.03]][[-8.47, -8.47, -8.47]][[0.02, -0.35]]
std:[[0.0, 0.0, 0.01]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.01, 0.0]]
MSE:[[0.03, 0.02, 0.03]][[0.03, 0.03, 0.03]][[8.47, 8.47, 8.47]][[0.02, 0.35]]
MSE(-DR):[[0.0, -0.01, 0.0]][[0.0, 0.0, 0.0]][[8.44, 8.44, 8.44]][[-0.01, 0.32]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
MC-based ATE = 0.04
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.31, 0.31, 0.28]][[0.39, 0.39, 0.39]][[-0.04, -0.04, -0.04]][0.28]
std:[[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.01]
MSE:[[0.31, 0.31, 0.28]][[0.39, 0.39, 0.39]][[0.04, 0.04, 0.04]][0.28]
MSE(-DR):[[0.0, 0.0, -0.03]][[0.08, 0.08, 0.08]][[-0.27, -0.27, -0.27]][-0.03]
better than DR_NO_MARL
==============
O_threshold = 15
MC-based mean and std of average reward:[8.339e+00 4.000e-03]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.38, -0.38, -0.37]][[-0.55, -0.55, -0.55]][[-8.34, -8.34, -8.34]][[-0.37, -0.22]]
std:[[0.02, 0.02, 0.01]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.0, 0.0]]
MSE:[[0.38, 0.38, 0.37]][[0.55, 0.55, 0.55]][[8.34, 8.34, 8.34]][[0.37, 0.22]]
MSE(-DR):[[0.0, 0.0, -0.01]][[0.17, 0.17, 0.17]][[7.96, 7.96, 7.96]][[-0.01, -0.16]]
better than DR_NO_MARL
MC-based ATE = -0.09
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.1, -0.09, -0.12]][[-0.2, -0.19, -0.2]][[0.09, 0.09, 0.09]][-0.11]
std:[[0.01, 0.01, 0.0]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.0]
MSE:[[0.1, 0.09, 0.12]][[0.2, 0.19, 0.2]][[0.09, 0.09, 0.09]][0.11]
MSE(-DR):[[0.0, -0.01, 0.02]][[0.1, 0.09, 0.1]][[-0.01, -0.01, -0.01]][0.01]
***** BETTER THAN [IS, DR_NO_MARL] *****
==============
time spent until now: 12.6 mins


---------------------------------------
[pattern_seed, T, sd_R] = [2, 672, 2]

max(u_O) =  27.57427313220561
O_threshold = 12
means of Order:

9.331 10.778 4.69 21.245 5.38

7.872 13.479 6.699 7.22 7.663

13.744 27.574 11.208 7.049 13.676

8.684 10.939 17.637 8.173 11.063

7.758 10.355 12.215 7.422 9.626

target policy:

0 0 0 1 0

0 1 0 0 0

1 1 0 0 1

0 0 1 0 0

0 0 1 0 0

number of reward locations:  7
O_threshold = 9
target policy:

1 1 0 1 0

0 1 0 0 0

1 1 1 0 1

0 1 1 0 1

0 1 1 0 1

number of reward locations:  14
O_threshold = 15
target policy:

0 0 0 1 0

0 0 0 0 0

0 1 0 0 0

0 0 1 0 0
```