```
Last login: Sun Mar 29 22:51:23 on ttys000
Run-Mac:~ mac$ cd ~/.ssh
Run-Mac:.ssh mac$ ssh -i "Runzhe.pem" ubuntu@ec2-3-228-4-227.compute-1.amazonaws.com
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 03:02:16 UTC 2020

  System load:  15.54              Processes:           211
  Usage of /:   56.8% of 15.45GB   Users logged in:     0
  Memory usage: 1%                 IP address for ens5: 172.31.6.17
  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

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


*** System restart required ***
Last login: Mon Mar 30 02:51:32 2020 from 107.13.161.147
ubuntu@ip-172-31-6-17:~$ export openblas_num_threads=1; export OMP_NUM_THREADS=1
ubuntu@ip-172-31-6-17:~$ python EC2.py
23:02, 03/29; num of cores:16

Basic setting:[sd_O, sd_D, sd_R, sd_u_O, w_O, w_A, lam] = [1, 1, 1, 0.4, 1, 1, 0.0001]


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

max(u_O) =  27.3
O_threshold = 12
means of Order:

22.3 12.9 16.3 27.0 23.3

7.5 16.1 10.4 10.6 13.0

11.7 19.7 14.9 11.6 13.2

12.6 20.0 10.2 12.5 7.8

4.0 14.3 15.6 8.2 27.3

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 = 10
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 = 14
target policy:

1 0 1 1 1

0 1 0 0 0

0 1 1 0 0

0 1 0 0 0

0 1 1 0 1

number of reward locations:  11
1 2 3 1 2 3
----------------------------------------
Value of Behaviour policy:9.554
O_threshold = 12
MC for this TARGET:[10.471, 0.009]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.03, -0.04, -0.09]][[0.02, 0.0, -0.0]][[-10.47, -10.47, -10.47]][[-0.1, -0.92]]
std:[[0.04, 0.04, 0.03]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.02, 0.01]]
MSE:[[0.05, 0.06, 0.09]][[0.02, 0.0, 0.0]][[10.47, 10.47, 10.47]][[0.1, 0.92]]
MSE(-DR):[[0.0, 0.01, 0.04]][[-0.03, -0.05, -0.05]][[10.42, 10.42, 10.42]][[0.05, 0.87]]
==============
O_threshold = 10
MC for this TARGET:[10.033, 0.009]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.19, 0.18, 0.13]][[0.39, 0.37, 0.38]][[-10.03, -10.03, -10.03]][[0.11, -0.48]]
std:[[0.07, 0.07, 0.0]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.0, 0.01]]
MSE:[[0.2, 0.19, 0.13]][[0.39, 0.37, 0.38]][[10.03, 10.03, 10.03]][[0.11, 0.48]]
MSE(-DR):[[0.0, -0.01, -0.07]][[0.19, 0.17, 0.18]][[9.83, 9.83, 9.83]][[-0.09, 0.28]]
better than DR_NO_MARL
MC-based ATE = -0.44
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.22, 0.22, 0.22]][[0.36, 0.37, 0.38]][[0.44, 0.44, 0.44]][0.22]
std:[[0.03, 0.03, 0.03]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.03]
MSE:[[0.22, 0.22, 0.22]][[0.36, 0.37, 0.38]][[0.44, 0.44, 0.44]][0.22]
MSE(-DR):[[0.0, 0.0, 0.0]][[0.14, 0.15, 0.16]][[0.22, 0.22, 0.22]][0.0]
***** BETTER THAN [IS, DR_NO_MARL] *****
==============
O_threshold = 14
MC for this TARGET:[10.463, 0.008]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.31, -0.32, -0.32]][[-0.32, -0.33, -0.35]][[-10.46, -10.46, -10.46]][[-0.33, -0.91]]
std:[[0.01, 0.01, 0.0]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.0, 0.01]]
MSE:[[0.31, 0.32, 0.32]][[0.32, 0.33, 0.35]][[10.46, 10.46, 10.46]][[0.33, 0.91]]
MSE(-DR):[[0.0, 0.01, 0.01]][[0.01, 0.02, 0.04]][[10.15, 10.15, 10.15]][[0.02, 0.6]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
MC-based ATE = -0.01
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.28, -0.28, -0.23]][[-0.34, -0.34, -0.35]][[0.01, 0.01, 0.01]][-0.23]
std:[[0.05, 0.04, 0.02]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.02]
MSE:[[0.28, 0.28, 0.23]][[0.34, 0.34, 0.35]][[0.01, 0.01, 0.01]][0.23]
MSE(-DR):[[0.0, 0.0, -0.05]][[0.06, 0.06, 0.07]][[-0.27, -0.27, -0.27]][-0.05]
better than DR_NO_MARL
==============
time spent until now: 2.4 mins


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

max(u_O) =  22.2
O_threshold = 12
means of Order:

21.1 8.6 8.9 7.2 15.6
```

```
4.4 22.2 8.1 12.5 10.0

19.8 4.8 9.7 9.5 17.3

7.1 10.3 7.8 11.2 13.9

7.1 17.4 15.8 13.5 15.8

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 = 10
target policy:

1 0 0 0 1

0 1 0 1 0

1 0 0 0 1

0 1 0 1 1

0 1 1 1 1

number of reward locations:  13
O_threshold = 14
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
--------------------------------------
Value of Behaviour policy:7.374
O_threshold = 12
MC for this TARGET:[8.188, 0.009]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.25, -0.27, -0.24]][[-0.33, -0.35, -0.35]][[-8.19, -8.19, -8.19]][[-0.25, -0.81]]
std:[[0.04, 0.04, 0.04]][[0.02, 0.02, 0.02]][[0.0, 0.0, 0.0]][[0.04, 0.0]]
MSE:[[0.25, 0.27, 0.24]][[0.33, 0.35, 0.35]][[8.19, 8.19, 8.19]][[0.25, 0.81]]
MSE(-DR):[[0.0, 0.02, -0.01]][[0.08, 0.1, 0.1]][[7.94, 7.94, 7.94]][[0.0, 0.56]]
better than DR_NO_MARL
==============
O_threshold = 10
MC for this TARGET:[8.046, 0.009]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.18, -0.2, -0.19]][[-0.14, -0.16, -0.16]][[-8.05, -8.05, -8.05]][[-0.21, -0.67]]
std:[[0.03, 0.03, 0.01]][[0.02, 0.02, 0.02]][[0.0, 0.0, 0.0]][[0.01, 0.0]]
MSE:[[0.18, 0.2, 0.19]][[0.14, 0.16, 0.16]][[8.05, 8.05, 8.05]][[0.21, 0.67]]
MSE(-DR):[[0.0, 0.02, 0.01]][[-0.04, -0.02, -0.02]][[7.87, 7.87, 7.87]][[0.03, 0.49]]
MC-based ATE = -0.14
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.07, 0.07, 0.04]][[0.19, 0.19, 0.19]][[0.14, 0.14, 0.14]][0.04]
std:[[0.02, 0.02, 0.03]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.03]
MSE:[[0.07, 0.07, 0.05]][[0.19, 0.19, 0.19]][[0.14, 0.14, 0.14]][0.05]
MSE(-DR):[[0.0, 0.0, -0.02]][[0.12, 0.12, 0.12]][[0.07, 0.07, 0.07]][-0.02]
better than DR_NO_MARL
==============
O_threshold = 14
MC for this TARGET:[8.165, 0.009]
```

```
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.44, -0.44, -0.38]][[-0.6, -0.62, -0.62]][[-8.16, -8.16, -8.16]][[-0.38, -0.79]]
std:[[0.02, 0.01, 0.0]][[0.03, 0.03, 0.02]][[0.0, 0.0, 0.0]][[0.0, 0.0]]
MSE:[[0.44, 0.44, 0.38]][[0.6, 0.62, 0.62]][[8.16, 8.16, 8.16]][[0.38, 0.79]]
MSE(-DR):[[0.0, 0.0, -0.06]][[0.16, 0.18, 0.18]][[7.72, 7.72, 7.72]][[-0.06, 0.35]]
better than DR_NO_MARL
MC-based ATE = -0.02
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.18, -0.18, -0.14]][[-0.27, -0.27, -0.27]][[0.02, 0.02, 0.02]][-0.13]
std:[[0.06, 0.06, 0.04]][[0.0, 0.01, 0.0]][[0.0, 0.0, 0.0]][0.04]
MSE:[[0.19, 0.19, 0.15]][[0.27, 0.27, 0.27]][[0.02, 0.02, 0.02]][0.14]
MSE(-DR):[[0.0, 0.0, -0.04]][[0.08, 0.08, 0.08]][[-0.17, -0.17, -0.17]][-0.05]
better than DR_NO_MARL
==============
time spent until now: 4.8 mins


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

max(u_O) =  27.6
O_threshold = 12
means of Order:

9.3 10.8 4.7 21.2 5.4

7.9 13.5 6.7 7.2 7.7

13.7 27.6 11.2 7.0 13.7

8.7 10.9 17.6 8.2 11.1

7.8 10.4 12.2 7.4 9.6

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 = 10
target policy:

0 1 0 1 0

0 1 0 0 0

1 1 1 0 1

0 1 1 0 1

0 1 1 0 0

number of reward locations:  12
O_threshold = 14
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
---------------------------------------
Value of Behaviour policy:6.891
```

```
O_threshold = 12
MC for this TARGET:[7.364, 0.008]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.26, -0.26, -0.27]][[-0.45, -0.46, -0.46]][[-7.36, -7.36, -7.36]][[-0.27, -0.47]]
std:[[0.04, 0.04, 0.01]][[0.03, 0.03, 0.02]][[0.0, 0.0, 0.0]][[0.0, 0.0]]
MSE:[[0.26, 0.26, 0.27]][[0.45, 0.46, 0.46]][[7.36, 7.36, 7.36]][[0.27, 0.47]]
MSE(-DR):[[0.0, 0.0, 0.01]][[0.19, 0.2, 0.2]][[7.1, 7.1, 7.1]][[0.01, 0.21]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
==============
O_threshold = 10
MC for this TARGET:[7.47, 0.008]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.09, -0.1, -0.09]][[-0.1, -0.12, -0.12]][[-7.47, -7.47, -7.47]][[-0.1, -0.58]]
std:[[0.04, 0.04, 0.02]][[0.01, 0.01, 0.0]][[0.0, 0.0, 0.0]][[0.02, 0.0]]
MSE:[[0.1, 0.11, 0.09]][[0.1, 0.12, 0.12]][[7.47, 7.47, 7.47]][[0.1, 0.58]]
MSE(-DR):[[0.0, 0.01, -0.01]][[0.0, 0.02, 0.02]][[7.37, 7.37, 7.37]][[0.0, 0.48]]
better than DR_NO_MARL
MC-based ATE = 0.11
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.17, 0.16, 0.18]][[0.35, 0.34, 0.34]][[-0.11, -0.11, -0.11]][0.17]
std:[[0.08, 0.08, 0.02]][[0.02, 0.02, 0.02]][[0.0, 0.0, 0.0]][0.02]
MSE:[[0.19, 0.18, 0.18]][[0.35, 0.34, 0.34]][[0.11, 0.11, 0.11]][0.17]
MSE(-DR):[[0.0, -0.01, -0.01]][[0.16, 0.15, 0.15]][[-0.08, -0.08, -0.08]][-0.02]
better than DR_NO_MARL
==============
O_threshold = 14
MC for this TARGET:[7.217, 0.008]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.34, -0.34, -0.31]][[-0.65, -0.66, -0.66]][[-7.22, -7.22, -7.22]][[-0.31, -0.33]]
std:[[0.05, 0.04, 0.03]][[0.02, 0.02, 0.01]][[0.0, 0.0, 0.0]][[0.02, 0.0]]
MSE:[[0.34, 0.34, 0.31]][[0.65, 0.66, 0.66]][[7.22, 7.22, 7.22]][[0.31, 0.33]]
MSE(-DR):[[0.0, 0.0, -0.03]][[0.31, 0.32, 0.32]][[6.88, 6.88, 6.88]][[-0.03, -0.01]]
better than DR_NO_MARL
MC-based ATE = -0.15
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.08, -0.08, -0.05]][[-0.2, -0.2, -0.19]][[0.15, 0.15, 0.15]][-0.04]
std:[[0.01, 0.01, 0.02]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][0.02]
MSE:[[0.08, 0.08, 0.05]][[0.2, 0.2, 0.19]][[0.15, 0.15, 0.15]][0.04]
MSE(-DR):[[0.0, 0.0, -0.03]][[0.12, 0.12, 0.11]][[0.07, 0.07, 0.07]][-0.04]
better than DR_NO_MARL
==============
time spent until now: 7.2 mins


---------------------------------------
[pattern_seed, T, sd_R] = [3, 672, 1]

max(u_O) =  22.5
O_threshold = 12
means of Order:

22.5 13.1 11.5 5.2 9.9

9.6 10.7 8.6 10.8 9.1

6.5 15.7 15.7 21.8 11.2

9.4 8.9 5.9 16.3 7.1

6.9 10.2 20.0 12.1 7.3

target policy:

1 1 0 0 0

0 0 0 0 0

0 1 1 1 0

0 0 0 1 0

0 0 1 1 0

number of reward locations:  8
O_threshold = 10
target policy:

1 1 1 0 0
```

```
0 1 0 1 0

0 1 1 1 1

0 0 0 1 0

0 1 1 1 0

number of reward locations:  13
O_threshold = 14
target policy:

1 0 0 0 0

0 0 0 0 0

0 1 1 1 0

0 0 0 1 0

0 0 1 0 0

number of reward locations:  6
1 2 3 1 2 3
--------------------------------------
Value of Behaviour policy:7.408
O_threshold = 12
MC for this TARGET:[8.015, 0.008]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.29, -0.29, -0.29]][[-0.46, -0.47, -0.47]][[-8.02, -8.02, -8.02]][[-0.29, -0.61]]
std:[[0.04, 0.04, 0.01]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.01, 0.01]]
MSE:[[0.29, 0.29, 0.29]][[0.46, 0.47, 0.47]][[8.02, 8.02, 8.02]][[0.29, 0.61]]
MSE(-DR):[[0.0, 0.0, 0.0]][[0.17, 0.18, 0.18]][[7.73, 7.73, 7.73]][[0.0, 0.32]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
==============
O_threshold = 10
MC for this TARGET:[7.939, 0.008]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.04, -0.04, -0.05]][[-0.02, -0.03, -0.04]][[-7.94, -7.94, -7.94]][[-0.06, -0.53]]
std:[[0.01, 0.01, 0.0]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.0, 0.01]]
MSE:[[0.04, 0.04, 0.05]][[0.02, 0.03, 0.04]][[7.94, 7.94, 7.94]][[0.06, 0.53]]
MSE(-DR):[[0.0, 0.0, 0.01]][[-0.02, -0.01, 0.0]][[7.9, 7.9, 7.9]][[0.02, 0.49]]
MC-based ATE = -0.08
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.25, 0.25, 0.23]][[0.44, 0.43, 0.43]][[0.08, 0.08, 0.08]][0.23]
std:[[0.03, 0.04, 0.01]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.02]
MSE:[[0.25, 0.25, 0.23]][[0.44, 0.43, 0.43]][[0.08, 0.08, 0.08]][0.23]
MSE(-DR):[[0.0, 0.0, -0.02]][[0.19, 0.18, 0.18]][[-0.17, -0.17, -0.17]][-0.02]
better than DR_NO_MARL
==============
O_threshold = 14
MC for this TARGET:[7.955, 0.008]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.36, -0.37, -0.36]][[-0.58, -0.58, -0.59]][[-7.96, -7.96, -7.96]][[-0.36, -0.55]]
std:[[0.05, 0.05, 0.02]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.02, 0.01]]
MSE:[[0.36, 0.37, 0.36]][[0.58, 0.58, 0.59]][[7.96, 7.96, 7.96]][[0.36, 0.55]]
MSE(-DR):[[0.0, 0.01, 0.0]][[0.22, 0.22, 0.23]][[7.6, 7.6, 7.6]][[0.0, 0.19]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
MC-based ATE = -0.06
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.07, -0.07, -0.07]][[-0.12, -0.11, -0.12]][[0.06, 0.06, 0.06]][-0.08]
std:[[0.01, 0.0, 0.01]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][0.01]
MSE:[[0.07, 0.07, 0.07]][[0.12, 0.11, 0.12]][[0.06, 0.06, 0.06]][0.08]
MSE(-DR):[[0.0, 0.0, 0.0]][[0.05, 0.04, 0.05]][[-0.01, -0.01, -0.01]][0.01]
***** BETTER THAN [IS, DR_NO_MARL] *****
==============
time spent until now: 9.6 mins


--------------------------------------
[pattern_seed, T, sd_R] = [4, 672, 1]

max(u_O) =  26.8
O_threshold = 12
means of Order:

11.2 13.5 7.4 14.5 9.3
```

```
5.8 8.5 14.0 12.6 7.0

14.1 10.6 13.1 12.6 6.9

12.7 8.6 20.5 14.7 11.2

7.4 11.3 11.8 6.8 26.8

target policy:

0 1 0 1 0

0 0 1 1 0

1 0 1 1 0

1 0 1 1 0

0 0 0 0 1

number of reward locations:  11
O_threshold = 10
target policy:

1 1 0 1 0

0 0 1 1 0

1 1 1 1 0

1 0 1 1 1

0 1 1 0 1

number of reward locations:  16
O_threshold = 14
target policy:

0 0 0 1 0

0 0 1 0 0

1 0 0 0 0

0 0 1 1 0

0 0 0 0 1

number of reward locations:  6
1 2 3 1 2 3
----------------------------------------
Value of Behaviour policy:7.806
O_threshold = 12
MC for this TARGET:[8.427, 0.008]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.15, -0.16, -0.17]][[-0.21, -0.22, -0.23]][[-8.43, -8.43, -8.43]][[-0.18, -0.62]]
std:[[0.03, 0.02, 0.03]][[0.02, 0.02, 0.02]][[0.0, 0.0, 0.0]][[0.02, 0.0]]
MSE:[[0.15, 0.16, 0.17]][[0.21, 0.22, 0.23]][[8.43, 8.43, 8.43]][[0.18, 0.62]]
MSE(-DR):[[0.0, 0.01, 0.02]][[0.06, 0.07, 0.08]][[8.28, 8.28, 8.28]][[0.03, 0.47]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
==============
O_threshold = 10
MC for this TARGET:[8.492, 0.008]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.02, -0.03, -0.09]][[0.12, 0.1, 0.1]][[-8.49, -8.49, -8.49]][[-0.1, -0.69]]
std:[[0.0, 0.0, 0.0]][[0.02, 0.02, 0.02]][[0.0, 0.0, 0.0]][[0.0, 0.0]]
MSE:[[0.02, 0.03, 0.09]][[0.12, 0.1, 0.1]][[8.49, 8.49, 8.49]][[0.1, 0.69]]
MSE(-DR):[[0.0, 0.01, 0.07]][[0.1, 0.08, 0.08]][[8.47, 8.47, 8.47]][[0.08, 0.67]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
MC-based ATE = 0.07
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.13, 0.13, 0.08]][[0.33, 0.33, 0.33]][[-0.07, -0.07, -0.07]][0.08]
std:[[0.02, 0.02, 0.03]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.02]
MSE:[[0.13, 0.13, 0.09]][[0.33, 0.33, 0.33]][[0.07, 0.07, 0.07]][0.08]
MSE(-DR):[[0.0, 0.0, -0.04]][[0.2, 0.2, 0.2]][[-0.06, -0.06, -0.06]][-0.05]
better than DR_NO_MARL
==============
```

```
O_threshold = 14
MC for this TARGET:[8.253, 0.008]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.35, -0.34, -0.33]][[-0.59, -0.59, -0.6]][[-8.25, -8.25, -8.25]][[-0.33, -0.45]]
std:[[0.07, 0.07, 0.05]][[0.03, 0.03, 0.03]][[0.0, 0.0, 0.0]][[0.04, 0.0]]
MSE:[[0.36, 0.35, 0.33]][[0.59, 0.59, 0.6]][[8.25, 8.25, 8.25]][[0.33, 0.45]]
MSE(-DR):[[0.0, -0.01, -0.03]][[0.23, 0.23, 0.24]][[7.89, 7.89, 7.89]][[-0.03, 0.09]]
better than DR_NO_MARL
MC-based ATE = -0.17
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.2, -0.19, -0.16]][[-0.38, -0.36, -0.37]][[0.17, 0.17, 0.17]][-0.15]
std:[[0.05, 0.05, 0.02]][[0.01, 0.01, 0.0]][[0.0, 0.0, 0.0]][0.02]
MSE:[[0.21, 0.2, 0.16]][[0.38, 0.36, 0.37]][[0.17, 0.17, 0.17]][0.15]
MSE(-DR):[[0.0, -0.01, -0.05]][[0.17, 0.15, 0.16]][[-0.04, -0.04, -0.04]][-0.06]
better than DR_NO_MARL
==============
time spent until now: 11.9 mins


--------------------------------------
[pattern_seed, T, sd_R] = [5, 672, 1]

max(u_O) =  29.1
O_threshold = 12
means of Order:

13.2 9.7 29.1 10.0 11.5

20.8 7.7 8.7 11.9 9.7

6.8 10.2 9.5 14.0 5.7

8.3 17.5 23.2 6.0 14.3

7.4 7.8 7.8 9.3 16.4

target policy:

1 0 1 0 0

1 0 0 0 0

0 0 0 1 0

0 1 1 0 1

0 0 0 0 1

number of reward locations:  8
O_threshold = 10
target policy:

1 0 1 0 1

1 0 0 1 0

0 1 0 1 0

0 1 1 0 1

0 0 0 0 1

number of reward locations:  11
O_threshold = 14
target policy:

0 0 1 0 0

1 0 0 0 0

0 0 0 1 0

0 1 1 0 1

0 0 0 0 1

number of reward locations:  7
1 2 3 1 2 3
```

```
---------------------------------------
Value of Behaviour policy:7.523
O_threshold = 12
MC for this TARGET:[8.307, 0.008]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.31, -0.32, -0.3]][[-0.53, -0.55, -0.55]][[-8.31, -8.31, -8.31]][[-0.31, -0.78]]
std:[[0.0, 0.0, 0.01]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.01, 0.0]]
MSE:[[0.31, 0.32, 0.3]][[0.53, 0.55, 0.55]][[8.31, 8.31, 8.31]][[0.31, 0.78]]
MSE(-DR):[[0.0, 0.01, -0.01]][[0.22, 0.24, 0.24]][[8.0, 8.0, 8.0]][[0.0, 0.47]]
better than DR_NO_MARL
==============
O_threshold = 10
MC for this TARGET:[8.19, 0.009]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.2, -0.21, -0.2]][[-0.26, -0.27, -0.27]][[-8.19, -8.19, -8.19]][[-0.21, -0.67]]
std:[[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.01, 0.0]]
MSE:[[0.2, 0.21, 0.2]][[0.26, 0.27, 0.27]][[8.19, 8.19, 8.19]][[0.21, 0.67]]
MSE(-DR):[[0.0, 0.01, 0.0]][[0.06, 0.07, 0.07]][[7.99, 7.99, 7.99]][[0.01, 0.47]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
MC-based ATE = -0.12
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.1, 0.1, 0.1]][[0.28, 0.28, 0.28]][[0.12, 0.12, 0.12]][0.1]
std:[[0.0, 0.0, 0.01]][[0.0, 0.0, 0.01]][[0.0, 0.0, 0.0]][0.01]
MSE:[[0.1, 0.1, 0.1]][[0.28, 0.28, 0.28]][[0.12, 0.12, 0.12]][0.1]
MSE(-DR):[[0.0, 0.0, 0.0]][[0.18, 0.18, 0.18]][[0.02, 0.02, 0.02]][0.0]
***** BETTER THAN [IS, DR_NO_MARL] *****
==============
O_threshold = 14
MC for this TARGET:[8.272, 0.008]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.34, -0.35, -0.33]][[-0.56, -0.57, -0.58]][[-8.27, -8.27, -8.27]][[-0.34, -0.75]]
std:[[0.0, 0.0, 0.02]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.02, 0.0]]
MSE:[[0.34, 0.35, 0.33]][[0.56, 0.57, 0.58]][[8.27, 8.27, 8.27]][[0.34, 0.75]]
MSE(-DR):[[0.0, 0.01, -0.01]][[0.22, 0.23, 0.24]][[7.93, 7.93, 7.93]][[0.0, 0.41]]
better than DR_NO_MARL
MC-based ATE = -0.04
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.04, -0.03, -0.03]][[-0.03, -0.03, -0.03]][[0.04, 0.04, 0.04]][-0.03]
std:[[0.0, 0.0, 0.01]][[0.01, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.01]
MSE:[[0.04, 0.03, 0.03]][[0.03, 0.03, 0.03]][[0.04, 0.04, 0.04]][0.03]
MSE(-DR):[[0.0, -0.01, -0.01]][[-0.01, -0.01, -0.01]][[0.0, 0.0, 0.0]][-0.01]
==============
time spent until now: 14.3 mins


---------------------------------------
[pattern_seed, T, sd_R] = [6, 672, 1]

max(u_O) =  31.6
O_threshold = 12
means of Order:

9.7 14.8 12.0 7.7 4.1

15.9 17.3 6.0 21.2 9.3

31.6 14.0 9.6 18.1 11.5

11.6 11.4 10.4 14.2 15.2

12.7 22.8 6.4 9.2 15.3

target policy:

0 1 1 0 0

1 1 0 1 0

1 1 0 1 0

0 0 0 1 1

1 1 0 0 1

number of reward locations:  13
O_threshold = 10
target policy:
```

```
0 1 1 0 0

1 1 0 1 0

1 1 0 1 1

1 1 1 1 1

1 1 0 0 1

number of reward locations:  17
O_threshold = 14
target policy:

0 1 0 0 0

1 1 0 1 0

1 1 0 1 0

0 0 0 1 1

0 1 0 0 1

number of reward locations:  11
```