```
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
```

```
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:[8.425 0.015]
Value of Behaviour policy:8.12
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.29, -0.29, -0.26]][[-0.33, -0.34, -0.33]][[-8.43, -8.43, -8.43]][[-0.26, -0.3]]
std:[[0.0, 0.0, 0.02]][[0.02, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.02, 0.0]]
MSE:[[0.29, 0.29, 0.26]][[0.33, 0.34, 0.33]][[8.43, 8.43, 8.43]][[0.26, 0.3]]
MSE(-DR):[[0.0, 0.0, -0.03]][[0.04, 0.05, 0.04]][[8.14, 8.14, 8.14]][[-0.03, 0.01]]
better than DR_NO_MARL
==============
O_threshold = 9
MC-based mean and std of average reward:[8.466 0.015]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.05, 0.04, 0.05]][[0.05, 0.04, 0.06]][[-8.47, -8.47, -8.47]][[0.04, -0.35]]
std:[[0.09, 0.08, 0.06]][[0.01, 0.01, 0.0]][[0.0, 0.0, 0.0]][[0.06, 0.0]]
MSE:[[0.1, 0.09, 0.08]][[0.05, 0.04, 0.06]][[8.47, 8.47, 8.47]][[0.07, 0.35]]
MSE(-DR):[[0.0, -0.01, -0.02]][[-0.05, -0.06, -0.04]][[8.37, 8.37, 8.37]][[-0.03, 0.25]]
MC-based ATE = 0.04
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.34, 0.34, 0.3]][[0.39, 0.38, 0.39]][[-0.04, -0.04, -0.04]][0.3]
std:[[0.09, 0.09, 0.04]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][0.04]
MSE:[[0.35, 0.35, 0.3]][[0.39, 0.38, 0.39]][[0.04, 0.04, 0.04]][0.3]
MSE(-DR):[[0.0, 0.0, -0.05]][[0.04, 0.03, 0.04]][[-0.31, -0.31, -0.31]][-0.05]
better than DR_NO_MARL
==============
O_threshold = 15
MC-based mean and std of average reward:[8.338 0.015]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.41, -0.4, -0.43]][[-0.54, -0.54, -0.53]][[-8.34, -8.34, -8.34]][[-0.42, -0.22]]
std:[[0.04, 0.04, 0.03]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.03, 0.0]]
MSE:[[0.41, 0.4, 0.43]][[0.54, 0.54, 0.53]][[8.34, 8.34, 8.34]][[0.42, 0.22]]
MSE(-DR):[[0.0, -0.01, 0.02]][[0.13, 0.13, 0.12]][[7.93, 7.93, 7.93]][[0.01, -0.19]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
MC-based ATE = -0.09
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.11, -0.1, -0.17]][[-0.2, -0.2, -0.2]][[0.09, 0.09, 0.09]][-0.16]
std:[[0.04, 0.04, 0.05]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.04]
MSE:[[0.12, 0.11, 0.18]][[0.2, 0.2, 0.2]][[0.09, 0.09, 0.09]][0.16]
MSE(-DR):[[0.0, -0.01, 0.06]][[0.08, 0.08, 0.08]][[-0.03, -0.03, -0.03]][0.04]
***** BETTER THAN [IS, DR_NO_MARL] *****
==============
time spent until now: 15.0 mins


--------------------------------------
[pattern_seed, T, sd_R] = [3, 672, 0]

max(u_O) =  22.5436040004391
O_threshold = 12
means of Order:

22.544 13.126 11.457 5.231 9.866

9.565 10.664 8.578 10.832 9.108

6.517 15.703 15.682 21.842 11.246

9.376 8.863 5.938 16.329 7.096

6.862 10.153 19.975 12.118 7.319

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 = 9
target policy:

1 1 1 0 1

1 1 0 1 1

0 1 1 1 1
```

```
1 0 0 1 0

0 1 1 1 0

number of reward locations:  17
O_threshold = 15
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
^CProcess Process-872:
Process Process-879:
Traceback (most recent call last):
Process Process-867:
Process Process-877:
  File "EC2.py", line 59, in <module>
Process Process-876:
Process Process-865:
Process Process-875:
Process Process-869:
Process Process-871:
Process Process-873:
Process Process-878:
Process Process-868:
Process Process-874:
Process Process-870:
    )
  File "/home/ubuntu/simu_funs.py", line 62, in simu
Process Process-866:
    value_reps = rep_seeds(once, OPE_rep_times)
  File "/home/ubuntu/_uti_basic.py", line 119, in rep_seeds
Process Process-880:
    return list(map(fun, range(rep_times)))
  File "/home/ubuntu/simu_funs.py", line 58, in once
    inner_parallel = inner_parallel)
  File "/home/ubuntu/simu_funs.py", line 185, in simu_once
    Ts = Ts, Ta = Ta, penalty = penalty, penalty_NMF = penalty_NMF,
  File "/home/ubuntu/main.py", line 130, in V_DR
    r = arr(parmap(getOneRegionValue, range(N), n_cores))
  File "/home/ubuntu/_uti_basic.py", line 74, in parmap
    sent = [q_in.put((i, x)) for i, x in enumerate(X)]
  File "/home/ubuntu/_uti_basic.py", line 74, in <listcomp>
    sent = [q_in.put((i, x)) for i, x in enumerate(X)]
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/queues.py", line 82, in put
    if not self._sem.acquire(block, timeout):
KeyboardInterrupt
ubuntu@ip-172-31-15-241:~$ python EC2.py
21:32, 03/29; num of cores:16

Basic setting:[sd_O, sd_D, sd_R, sd_u_O, w_O, w_A, lam] = [0, 0, 0, 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
----------------------------------------
Traceback (most recent call last):
  File "EC2.py", line 58, in <module>
    file = file, print_flag_target = False
  File "/home/ubuntu/simu_funs.py", line 79, in simu
    V_behav = np.mean(V_OPE, 0)[-1]
UnboundLocalError: local variable 'V_OPE' referenced before assignment
ubuntu@ip-172-31-15-241:~$ ^C
ubuntu@ip-172-31-15-241:~$ python EC2.py
21:36, 03/29; num of cores:16

Basic setting:[sd_O, sd_D, sd_R, sd_u_O, w_O, w_A, lam] = [0, 0, 0, 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
--------------------------------------
Value of Behaviour policy:11.346
O_threshold = 12
MC for this TARGET:[11.879, 0.0]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[nan, 0.15, nan]][[nan, 0.15, nan]][[-11.88, -11.88, -11.88]][[nan, -0.53]]
std:[[nan, 0.0, nan]][[nan, 0.0, nan]][[0.0, 0.0, 0.0]][[nan, 0.0]]
MSE:[[nan, 0.15, nan]][[nan, 0.15, nan]][[11.88, 11.88, 11.88]][[nan, 0.53]]
MSE(-DR):[[nan, nan, nan]][[nan, nan, nan]][[nan, nan, nan]][[nan, nan]]
==============
O_threshold = 9
MC for this TARGET:[11.634, 0.0]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[nan, 0.43, nan]][[nan, 0.45, nan]][[-11.63, -11.63, -11.63]][[nan, -0.29]]
std:[[nan, 0.01, nan]][[nan, 0.0, nan]][[0.0, 0.0, 0.0]][[nan, 0.0]]
MSE:[[nan, 0.43, nan]][[nan, 0.45, nan]][[11.63, 11.63, 11.63]][[nan, 0.29]]
MSE(-DR):[[nan, nan, nan]][[nan, nan, nan]][[nan, nan, nan]][[nan, nan]]
MC-based ATE = -0.24
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[nan, 0.29, nan]][[nan, 0.3, nan]][[0.24, 0.24, 0.24]][nan]
std:[[nan, 0.01, nan]][[nan, 0.0, nan]][[0.0, 0.0, 0.0]][nan]
MSE:[[nan, 0.29, nan]][[nan, 0.3, nan]][[0.24, 0.24, 0.24]][nan]
MSE(-DR):[[nan, nan, nan]][[nan, nan, nan]][[nan, nan, nan]][nan]
==============
O_threshold = 15
MC for this TARGET:[11.85, 0.0]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[nan, -0.26, nan]][[nan, -0.37, nan]][[-11.85, -11.85, -11.85]][[nan, -0.5]]
std:[[nan, 0.01, nan]][[nan, 0.0, nan]][[0.0, 0.0, 0.0]][[nan, 0.0]]
MSE:[[nan, 0.26, nan]][[nan, 0.37, nan]][[11.85, 11.85, 11.85]][[nan, 0.5]]
MSE(-DR):[[nan, nan, nan]][[nan, nan, nan]][[nan, nan, nan]][[nan, nan]]
MC-based ATE = -0.03
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[nan, -0.41, nan]][[nan, -0.52, nan]][[0.03, 0.03, 0.03]][nan]
std:[[nan, 0.01, nan]][[nan, 0.0, nan]][[0.0, 0.0, 0.0]][nan]
MSE:[[nan, 0.41, nan]][[nan, 0.52, nan]][[0.03, 0.03, 0.03]][nan]
MSE(-DR):[[nan, nan, nan]][[nan, nan, nan]][[nan, nan, nan]][nan]
==============
time spent until now: 2.4 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 ^CTraceback (most recent call last):
  File "EC2.py", line 58, in <module>
Process Process-173:
Process Process-166:
Process Process-170:
Process Process-167:
Process Process-171:
Process Process-163:
    file = file, print_flag_target = False
Process Process-168:
Process Process-175:
Process Process-161:
Process Process-174:
Process Process-172:
Process Process-176:
Process Process-164:
  File "/home/ubuntu/simu_funs.py", line 62, in simu
Process Process-162:
    value_reps = rep_seeds(once, OPE_rep_times)
  File "/home/ubuntu/_uti_basic.py", line 119, in rep_seeds
    return list(map(fun, range(rep_times)))
  File "/home/ubuntu/simu_funs.py", line 58, in once
    inner_parallel = inner_parallel)
  File "/home/ubuntu/simu_funs.py", line 190, in simu_once
    inner_parallel = inner_parallel)
  File "/home/ubuntu/main.py", line 130, in V_DR
    r = arr(parmap(getOneRegionValue, range(N), n_cores))
  File "/home/ubuntu/_uti_basic.py", line 74, in parmap
    sent = [q_in.put((i, x)) for i, x in enumerate(X)]
  File "/home/ubuntu/_uti_basic.py", line 74, in <listcomp>
    sent = [q_in.put((i, x)) for i, x in enumerate(X)]
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/queues.py", line 82, in put
    if not self._sem.acquire(block, timeout):
KeyboardInterrupt
ubuntu@ip-172-31-15-241:~$ python EC2.py
21:39, 03/29; num of cores:16

Basic setting:[sd_O, sd_D, sd_R, sd_u_O, w_O, w_A, lam] = [0.01, 0.01, 0.01, 0.4, 1, 1, 0.0001]


----------------------------------------
[pattern_seed, T, sd_R] = [0, 672, 0.01]

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
----------------------------------------
Value of Behaviour policy:11.343
O_threshold = 12
MC for this TARGET:[11.878, 0.0]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.15, 0.14, 0.12]][[0.16, 0.15, 0.16]][[-11.88, -11.88, -11.88]][[0.11, -0.54]]
std:[[0.01, 0.0, 0.0]][[0.01, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.0, 0.0]]
MSE:[[0.15, 0.14, 0.12]][[0.16, 0.15, 0.16]][[11.88, 11.88, 11.88]][[0.11, 0.54]]
MSE(-DR):[[0.0, -0.01, -0.03]][[0.01, 0.0, 0.01]][[11.73, 11.73, 11.73]][[-0.04, 0.39]]
better than DR_NO_MARL
==============
O_threshold = 9
MC for this TARGET:[11.633, 0.0]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.44, 0.43, 0.4]][[0.46, 0.46, 0.46]][[-11.63, -11.63, -11.63]][[0.4, -0.29]]
std:[[0.0, 0.0, 0.0]][[0.01, 0.0, 0.01]][[0.0, 0.0, 0.0]][[0.0, 0.0]]
MSE:[[0.44, 0.43, 0.4]][[0.46, 0.46, 0.46]][[11.63, 11.63, 11.63]][[0.4, 0.29]]
MSE(-DR):[[0.0, -0.01, -0.04]][[0.02, 0.02, 0.02]][[11.19, 11.19, 11.19]][[-0.04, -0.15]]
better than DR_NO_MARL
MC-based ATE = -0.25
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.29, 0.29, 0.28]][[0.3, 0.3, 0.31]][[0.25, 0.25, 0.25]][0.28]
std:[[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.0]
MSE:[[0.29, 0.29, 0.28]][[0.3, 0.3, 0.31]][[0.25, 0.25, 0.25]][0.28]
MSE(-DR):[[0.0, 0.0, -0.01]][[0.01, 0.01, 0.02]][[-0.04, -0.04, -0.04]][-0.01]
better than DR_NO_MARL
==============
O_threshold = 15
MC for this TARGET:[11.849, 0.0]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.26, -0.26, -0.21]][[-0.35, -0.36, -0.35]][[-11.85, -11.85, -11.85]][[-0.21, -0.51]]
std:[[0.01, 0.01, 0.01]][[0.01, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.0, 0.0]]
MSE:[[0.26, 0.26, 0.21]][[0.35, 0.36, 0.35]][[11.85, 11.85, 11.85]][[0.21, 0.51]]
MSE(-DR):[[0.0, 0.0, -0.05]][[0.09, 0.1, 0.09]][[11.59, 11.59, 11.59]][[-0.05, 0.25]]
better than DR_NO_MARL
MC-based ATE = -0.03
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.4, -0.4, -0.33]][[-0.51, -0.51, -0.51]][[0.03, 0.03, 0.03]][-0.33]
std:[[0.01, 0.02, 0.0]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.0]
MSE:[[0.4, 0.4, 0.33]][[0.51, 0.51, 0.51]][[0.03, 0.03, 0.03]][0.33]
MSE(-DR):[[0.0, 0.0, -0.07]][[0.11, 0.11, 0.11]][[-0.37, -0.37, -0.37]][-0.07]
better than DR_NO_MARL
==============
time spent until now: 2.4 mins


----------------------------------------
[pattern_seed, T, sd_R] = [1, 672, 0.01]

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 ^CProcess Process-207:
Process Process-195:
Process Process-204:
Process Process-205:
Process Process-194:
Process Process-198:
Traceback (most recent call last):
  File "EC2.py", line 62, in <module>
Process Process-201:
Process Process-202:
Process Process-200:
    file = file, print_flag_target = False
Process Process-208:
  File "/home/ubuntu/simu_funs.py", line 62, in simu
Process Process-203:
Process Process-197:
Traceback (most recent call last):
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 297, in _bootstrap
    self.run()
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 99, in run
    self._target(*self._args, **self._kwargs)
  File "/home/ubuntu/_uti_basic.py", line 62, in fun
    q_out.put((i, f(x)))
  File "/home/ubuntu/main.py", line 85, in getOneRegionValue
    spatial = False)
  File "/home/ubuntu/main.py", line 236, in getWeight
    epsilon = epsilon, spatial = spatial, mean_field = mean_field)
  File "/home/ubuntu/weight.py", line 297, in train
    self.policy_ratio2: policy_ratio2
KeyboardInterrupt
Process Process-196:
Process Process-193:
Process Process-206:
Traceback (most recent call last):
Traceback (most recent call last):
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 297, in _bootstrap
    self.run()
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 99, in run
    self._target(*self._args, **self._kwargs)
```

```
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 297, in _bootstrap
    self.run()
  File "/home/ubuntu/_uti_basic.py", line 62, in fun
    q_out.put((i, f(x)))
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 99, in run
    self._target(*self._args, **self._kwargs)
  File "/home/ubuntu/main.py", line 85, in getOneRegionValue
    spatial = False)
Traceback (most recent call last):
  File "/home/ubuntu/_uti_basic.py", line 62, in fun
    q_out.put((i, f(x)))
  File "/home/ubuntu/main.py", line 236, in getWeight
    epsilon = epsilon, spatial = spatial, mean_field = mean_field)
  File "/home/ubuntu/main.py", line 85, in getOneRegionValue
    spatial = False)
  File "/home/ubuntu/weight.py", line 297, in train
    self.policy_ratio2: policy_ratio2
Traceback (most recent call last):
Traceback (most recent call last):
  File "/home/ubuntu/main.py", line 236, in getWeight
    epsilon = epsilon, spatial = spatial, mean_field = mean_field)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 950, in run
    run_metadata_ptr)
  File "/home/ubuntu/weight.py", line 297, in train
    self.policy_ratio2: policy_ratio2
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1167, in _run
    final_fetches = fetch_handler.fetches()
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 950, in run
    run_metadata_ptr)
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 297, in _bootstrap
    self.run()
Traceback (most recent call last):
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1173, in _run
    feed_dict_tensor, options, run_metadata)
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 297, in _bootstrap
    self.run()
Traceback (most recent call last):
Traceback (most recent call last):
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 99, in run
    self._target(*self._args, **self._kwargs)
KeyboardInterrupt
Traceback (most recent call last):
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1350, in _do_run
    run_metadata)
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 297, in _bootstrap
    self.run()
Traceback (most recent call last):
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 99, in run
    self._target(*self._args, **self._kwargs)
  File "/home/ubuntu/_uti_basic.py", line 62, in fun
    q_out.put((i, f(x)))
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1356, in _do_call
    return fn(*args)
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 99, in run
    self._target(*self._args, **self._kwargs)
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 297, in _bootstrap
    self.run()
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 297, in _bootstrap
    self.run()
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 297, in _bootstrap
    self.run()
  File "/home/ubuntu/_uti_basic.py", line 62, in fun
    q_out.put((i, f(x)))
  File "/home/ubuntu/main.py", line 85, in getOneRegionValue
    spatial = False)
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 297, in _bootstrap
    self.run()
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1341, in _run_fn
    options, feed_dict, fetch_list, target_list, run_metadata)
  File "/home/ubuntu/_uti_basic.py", line 62, in fun
    q_out.put((i, f(x)))
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 297, in _bootstrap
    self.run()
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 99, in run
    self._target(*self._args, **self._kwargs)
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 99, in run
    self._target(*self._args, **self._kwargs)
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 99, in run
    self._target(*self._args, **self._kwargs)
  File "/home/ubuntu/main.py", line 236, in getWeight
    epsilon = epsilon, spatial = spatial, mean_field = mean_field)
  File "/home/ubuntu/main.py", line 85, in getOneRegionValue
    spatial = False)
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 99, in run
    self._target(*self._args, **self._kwargs)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1429, in _call_tf_sessionrun
    run_metadata)
  File "/home/ubuntu/main.py", line 85, in getOneRegionValue
    spatial = False)
```

```
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 99, in run
    self._target(*self._args, **self._kwargs)
  File "/home/ubuntu/main.py", line 236, in getWeight
    epsilon = epsilon, spatial = spatial, mean_field = mean_field)
  File "/home/ubuntu/_uti_basic.py", line 62, in fun
    q_out.put((i, f(x)))
  File "/home/ubuntu/_uti_basic.py", line 62, in fun
    q_out.put((i, f(x)))
  File "/home/ubuntu/_uti_basic.py", line 62, in fun
    q_out.put((i, f(x)))
  File "/home/ubuntu/weight.py", line 297, in train
    self.policy_ratio2: policy_ratio2
  File "/home/ubuntu/main.py", line 236, in getWeight
    epsilon = epsilon, spatial = spatial, mean_field = mean_field)
  File "/home/ubuntu/_uti_basic.py", line 62, in fun
    q_out.put((i, f(x)))
  File "/home/ubuntu/_uti_basic.py", line 62, in fun
    q_out.put((i, f(x)))
KeyboardInterrupt
  File "/home/ubuntu/weight.py", line 297, in train
    self.policy_ratio2: policy_ratio2
  File "/home/ubuntu/main.py", line 85, in getOneRegionValue
    spatial = False)
  File "/home/ubuntu/main.py", line 85, in getOneRegionValue
    spatial = False)
  File "/home/ubuntu/main.py", line 85, in getOneRegionValue
    spatial = False)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 950, in run
    run_metadata_ptr)
  File "/home/ubuntu/weight.py", line 297, in train
    self.policy_ratio2: policy_ratio2
  File "/home/ubuntu/main.py", line 85, in getOneRegionValue
    spatial = False)
  File "/home/ubuntu/main.py", line 85, in getOneRegionValue
    spatial = False)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 950, in run
    run_metadata_ptr)
  File "/home/ubuntu/main.py", line 236, in getWeight
    epsilon = epsilon, spatial = spatial, mean_field = mean_field)
  File "/home/ubuntu/main.py", line 236, in getWeight
    epsilon = epsilon, spatial = spatial, mean_field = mean_field)
  File "/home/ubuntu/main.py", line 236, in getWeight
    epsilon = epsilon, spatial = spatial, mean_field = mean_field)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1173, in _run
    feed_dict_tensor, options, run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 950, in run
    run_metadata_ptr)
  File "/home/ubuntu/main.py", line 236, in getWeight
    epsilon = epsilon, spatial = spatial, mean_field = mean_field)
  File "/home/ubuntu/main.py", line 236, in getWeight
    epsilon = epsilon, spatial = spatial, mean_field = mean_field)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1173, in _run
    feed_dict_tensor, options, run_metadata)
  File "/home/ubuntu/weight.py", line 297, in train
    self.policy_ratio2: policy_ratio2
  File "/home/ubuntu/weight.py", line 297, in train
    self.policy_ratio2: policy_ratio2
  File "/home/ubuntu/weight.py", line 297, in train
    self.policy_ratio2: policy_ratio2
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1350, in _do_run
    run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1173, in _run
    feed_dict_tensor, options, run_metadata)
  File "/home/ubuntu/weight.py", line 297, in train
    self.policy_ratio2: policy_ratio2
  File "/home/ubuntu/weight.py", line 297, in train
    self.policy_ratio2: policy_ratio2
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1350, in _do_run
    run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 950, in run
    run_metadata_ptr)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 950, in run
    run_metadata_ptr)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 950, in run
    run_metadata_ptr)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1356, in _do_call
    return fn(*args)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1334, in _do_run
    targets = [op._c_op for op in target_list]
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 950, in run
    run_metadata_ptr)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 950, in run
    run_metadata_ptr)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1356, in _do_call
    return fn(*args)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1173, in _run
    feed_dict_tensor, options, run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1173, in _run
    feed_dict_tensor, options, run_metadata)
```

```
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1173, in _run
    feed_dict_tensor, options, run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1341, in _run_fn
    options, feed_dict, fetch_list, target_list, run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1334, in <listcomp>
    targets = [op._c_op for op in target_list]
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1173, in _run
    feed_dict_tensor, options, run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1173, in _run
    feed_dict_tensor, options, run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1341, in _run_fn
    options, feed_dict, fetch_list, target_list, run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1350, in _do_run
    run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1350, in _do_run
    run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1350, in _do_run
    run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1429, in _call_tf_sessionrun
    run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1350, in _do_run
    run_metadata)
KeyboardInterrupt
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1350, in _do_run
    run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1429, in _call_tf_sessionrun
    run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1356, in _do_call
    return fn(*args)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1356, in _do_call
    return fn(*args)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1356, in _do_call
    return fn(*args)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1356, in _do_call
    return fn(*args)
KeyboardInterrupt
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1356, in _do_call
    return fn(*args)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1341, in _run_fn
    options, feed_dict, fetch_list, target_list, run_metadata)
KeyboardInterrupt
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1341, in _run_fn
    options, feed_dict, fetch_list, target_list, run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1341, in _run_fn
    options, feed_dict, fetch_list, target_list, run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1341, in _run_fn
    options, feed_dict, fetch_list, target_list, run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1341, in _run_fn
    options, feed_dict, fetch_list, target_list, run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1429, in _call_tf_sessionrun
    run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1429, in _call_tf_sessionrun
    run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1429, in _call_tf_sessionrun
    run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1429, in _call_tf_sessionrun
    run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1429, in _call_tf_sessionrun
    run_metadata)
KeyboardInterrupt
KeyboardInterrupt
KeyboardInterrupt
KeyboardInterrupt
KeyboardInterrupt
Process Process-199:
    value_reps = rep_seeds(once, OPE_rep_times)
  File "/home/ubuntu/_uti_basic.py", line 119, in rep_seeds
    return list(map(fun, range(rep_times)))
  File "/home/ubuntu/simu_funs.py", line 58, in once
    inner_parallel = inner_parallel)
  File "/home/ubuntu/simu_funs.py", line 190, in simu_once
    inner_parallel = inner_parallel)
  File "/home/ubuntu/main.py", line 130, in V_DR
Traceback (most recent call last):
    r = arr(parmap(getOneRegionValue, range(N), n_cores))
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 297, in _bootstrap
    self.run()
  File "/home/ubuntu/_uti_basic.py", line 74, in parmap
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 99, in run
    self._target(*self._args, **self._kwargs)
  File "/home/ubuntu/_uti_basic.py", line 62, in fun
    q_out.put((i, f(x)))
  File "/home/ubuntu/main.py", line 85, in getOneRegionValue
    spatial = False)
  File "/home/ubuntu/main.py", line 236, in getWeight
    epsilon = epsilon, spatial = spatial, mean_field = mean_field)
  File "/home/ubuntu/weight.py", line 297, in train
    self.policy_ratio2: policy_ratio2
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 950, in run
```

```
    run_metadata_ptr)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1173, in _run
    feed_dict_tensor, options, run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1350, in _do_run
    run_metadata)
    sent = [q_in.put((i, x)) for i, x in enumerate(X)]
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1356, in _do_call
    return fn(*args)
  File "/home/ubuntu/_uti_basic.py", line 74, in <listcomp>
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1341, in _run_fn
    options, feed_dict, fetch_list, target_list, run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1429, in _call_tf_sessionrun
    run_metadata)
KeyboardInterrupt
    sent = [q_in.put((i, x)) for i, x in enumerate(X)]
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/queues.py", line 82, in put
    if not self._sem.acquire(block, timeout):
KeyboardInterrupt
Traceback (most recent call last):
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 297, in _bootstrap
    self.run()
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 99, in run
    self._target(*self._args, **self._kwargs)
  File "/home/ubuntu/_uti_basic.py", line 62, in fun
    q_out.put((i, f(x)))
  File "/home/ubuntu/main.py", line 85, in getOneRegionValue
    spatial = False)
  File "/home/ubuntu/main.py", line 236, in getWeight
    epsilon = epsilon, spatial = spatial, mean_field = mean_field)
  File "/home/ubuntu/weight.py", line 297, in train
    self.policy_ratio2: policy_ratio2
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 950, in run
    run_metadata_ptr)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1173, in _run
    feed_dict_tensor, options, run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1350, in _do_run
    run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1356, in _do_call
    return fn(*args)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1339, in _run_fn
    self._extend_graph()
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1373, in _extend_graph
    with self._graph._session_run_lock():  # pylint: disable=protected-access
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/util/lock_util.py", line 124, in __enter__
    self._lock.acquire(self._group_id)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/util/lock_util.py", line 91, in acquire
    while self._another_group_active(group_id):
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/util/lock_util.py", line 108, in _another_group_active
    c > 0 for g, c in enumerate(self._group_member_counts) if g != group_id)
KeyboardInterrupt
ubuntu@ip-172-31-15-241:~$ python EC2.py
21:43, 03/29; num of cores:16

Basic setting:[sd_O, sd_D, sd_R, sd_u_O, w_O, w_A, lam] = [0.01, 0.01, 0.01, 0.4, 1, 1, 1e-05]


---------------------------------------
[pattern_seed, T, sd_R] = [0, 672, 0.01]

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
----------------------------------------
Value of Behaviour policy:11.343
O_threshold = 12
MC for this TARGET:[11.878, 0.0]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.14, 0.14, 0.11]][[0.16, 0.15, 0.15]][[-11.88, -11.88, -11.88]][[0.11, -0.54]]
std:[[0.01, 0.01, 0.0]][[0.01, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.0, 0.0]]
MSE:[[0.14, 0.14, 0.11]][[0.16, 0.15, 0.15]][[11.88, 11.88, 11.88]][[0.11, 0.54]]
MSE(-DR):[[0.0, 0.0, -0.03]][[0.02, 0.01, 0.01]][[11.74, 11.74, 11.74]][[-0.03, 0.4]]
better than DR_NO_MARL
==============
O_threshold = 9
MC for this TARGET:[11.633, 0.0]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.43, 0.43, 0.4]][[0.46, 0.46, 0.46]][[-11.63, -11.63, -11.63]][[0.39, -0.29]]
std:[[0.01, 0.01, 0.0]][[0.01, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.0, 0.0]]
MSE:[[0.43, 0.43, 0.4]][[0.46, 0.46, 0.46]][[11.63, 11.63, 11.63]][[0.39, 0.29]]
MSE(-DR):[[0.0, 0.0, -0.03]][[0.03, 0.03, 0.03]][[11.2, 11.2, 11.2]][[-0.04, -0.14]]
better than DR_NO_MARL
MC-based ATE = -0.25
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.29, 0.29, 0.28]][[0.3, 0.3, 0.31]][[0.25, 0.25, 0.25]][0.28]
std:[[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.0]
MSE:[[0.29, 0.29, 0.28]][[0.3, 0.3, 0.31]][[0.25, 0.25, 0.25]][0.28]
MSE(-DR):[[0.0, 0.0, -0.01]][[0.01, 0.01, 0.02]][[-0.04, -0.04, -0.04]][-0.01]
better than DR_NO_MARL
==============
O_threshold = 15
MC for this TARGET:[11.849, 0.0]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.25, -0.25, -0.21]][[-0.35, -0.36, -0.36]][[-11.85, -11.85, -11.85]][[-0.21, -0.51]]
std:[[0.02, 0.02, 0.0]][[0.01, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.0, 0.0]]
MSE:[[0.25, 0.25, 0.21]][[0.35, 0.36, 0.36]][[11.85, 11.85, 11.85]][[0.21, 0.51]]
MSE(-DR):[[0.0, 0.0, -0.04]][[0.1, 0.11, 0.11]][[11.6, 11.6, 11.6]][[-0.04, 0.26]]
better than DR_NO_MARL
MC-based ATE = -0.03
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.4, -0.4, -0.32]][[-0.51, -0.51, -0.51]][[0.03, 0.03, 0.03]][-0.32]
std:[[0.03, 0.03, 0.0]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.0]
MSE:[[0.4, 0.4, 0.32]][[0.51, 0.51, 0.51]][[0.03, 0.03, 0.03]][0.32]
MSE(-DR):[[0.0, 0.0, -0.08]][[0.11, 0.11, 0.11]][[-0.37, -0.37, -0.37]][-0.08]
better than DR_NO_MARL
==============
time spent until now: 5.4 mins


----------------------------------------
[pattern_seed, T, sd_R] = [1, 672, 0.01]

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 ^CTraceback (most recent call last):
  File "EC2.py", line 69, in <module>
Process Process-206:
Process Process-194:
    file = file, print_flag_target = False
  File "/home/ubuntu/simu_funs.py", line 62, in simu
    value_reps = rep_seeds(once, OPE_rep_times)
  File "/home/ubuntu/_uti_basic.py", line 119, in rep_seeds
Process Process-195:
    return list(map(fun, range(rep_times)))
  File "/home/ubuntu/simu_funs.py", line 58, in once
Process Process-203:
Process Process-197:
Process Process-201:
Process Process-200:
    inner_parallel = inner_parallel)
  File "/home/ubuntu/simu_funs.py", line 190, in simu_once
Process Process-207:
Process Process-204:
    inner_parallel = inner_parallel)
  File "/home/ubuntu/main.py", line 130, in V_DR
Process Process-208:
    r = arr(parmap(getOneRegionValue, range(N), n_cores))
  File "/home/ubuntu/_uti_basic.py", line 74, in parmap
    sent = [q_in.put((i, x)) for i, x in enumerate(X)]
  File "/home/ubuntu/_uti_basic.py", line 74, in <listcomp>
Process Process-196:
    sent = [q_in.put((i, x)) for i, x in enumerate(X)]
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/queues.py", line 82, in put
    if not self._sem.acquire(block, timeout):
KeyboardInterrupt
Process Process-198:
Process Process-202:
Traceback (most recent call last):
Process Process-205:
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 297, in _bootstrap
    self.run()
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 99, in run
    self._target(*self._args, **self._kwargs)
  File "/home/ubuntu/_uti_basic.py", line 62, in fun
    q_out.put((i, f(x)))
  File "/home/ubuntu/main.py", line 85, in getOneRegionValue
    spatial = False)
  File "/home/ubuntu/main.py", line 236, in getWeight
    epsilon = epsilon, spatial = spatial, mean_field = mean_field)
  File "/home/ubuntu/weight.py", line 297, in train
    self.policy_ratio2: policy_ratio2
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 950, in run
```

```
    run_metadata_ptr)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1176, in _run
    return fetch_handler.build_results(self, results)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 564, in build_results
    return self._fetch_mapper.build_results(full_values)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 383, in build_results
    results.append(m.build_results([values[j] for j in vi]))
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 319, in build_results
    def build_results(self, values):
Traceback (most recent call last):
Traceback (most recent call last):
KeyboardInterrupt
Traceback (most recent call last):
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 297, in _bootstrap
    self.run()
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 297, in _bootstrap
    self.run()
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 99, in run
    self._target(*self._args, **self._kwargs)
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 99, in run
    self._target(*self._args, **self._kwargs)
  File "/home/ubuntu/_uti_basic.py", line 62, in fun
    q_out.put((i, f(x)))
  File "/home/ubuntu/_uti_basic.py", line 62, in fun
    q_out.put((i, f(x)))
  File "/home/ubuntu/main.py", line 85, in getOneRegionValue
    spatial = False)
  File "/home/ubuntu/main.py", line 85, in getOneRegionValue
    spatial = False)
  File "/home/ubuntu/main.py", line 236, in getWeight
    epsilon = epsilon, spatial = spatial, mean_field = mean_field)
  File "/home/ubuntu/main.py", line 236, in getWeight
    epsilon = epsilon, spatial = spatial, mean_field = mean_field)
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 297, in _bootstrap
    self.run()
  File "/home/ubuntu/weight.py", line 297, in train
    self.policy_ratio2: policy_ratio2
  File "/home/ubuntu/weight.py", line 297, in train
    self.policy_ratio2: policy_ratio2
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 99, in run
    self._target(*self._args, **self._kwargs)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 950, in run
    run_metadata_ptr)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 950, in run
    run_metadata_ptr)
  File "/home/ubuntu/_uti_basic.py", line 62, in fun
    q_out.put((i, f(x)))
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1173, in _run
    feed_dict_tensor, options, run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1350, in _do_run
    run_metadata)
  File "/home/ubuntu/main.py", line 85, in getOneRegionValue
    spatial = False)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1173, in _run
    feed_dict_tensor, options, run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1356, in _do_call
    return fn(*args)
  File "/home/ubuntu/main.py", line 236, in getWeight
    epsilon = epsilon, spatial = spatial, mean_field = mean_field)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1350, in _do_run
    run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1341, in _run_fn
    options, feed_dict, fetch_list, target_list, run_metadata)
  File "/home/ubuntu/weight.py", line 297, in train
    self.policy_ratio2: policy_ratio2
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1356, in _do_call
    return fn(*args)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1429, in _call_tf_sessionrun
    run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 950, in run
    run_metadata_ptr)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1341, in _run_fn
    options, feed_dict, fetch_list, target_list, run_metadata)
KeyboardInterrupt
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1429, in _call_tf_sessionrun
    run_metadata)
KeyboardInterrupt
KeyboardInterrupt
Traceback (most recent call last):
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 297, in _bootstrap
    self.run()
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 99, in run
    self._target(*self._args, **self._kwargs)
  File "/home/ubuntu/_uti_basic.py", line 62, in fun
    q_out.put((i, f(x)))
  File "/home/ubuntu/main.py", line 85, in getOneRegionValue
    spatial = False)
  File "/home/ubuntu/main.py", line 236, in getWeight
    epsilon = epsilon, spatial = spatial, mean_field = mean_field)
```

```
  File "/home/ubuntu/weight.py", line 297, in train
    self.policy_ratio2: policy_ratio2
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 950, in run
    run_metadata_ptr)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1173, in _run
    feed_dict_tensor, options, run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1350, in _do_run
    run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1356, in _do_call
    return fn(*args)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1341, in _run_fn
    options, feed_dict, fetch_list, target_list, run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1429, in _call_tf_sessionrun
    run_metadata)
KeyboardInterrupt
Traceback (most recent call last):
Traceback (most recent call last):
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 297, in _bootstrap
    self.run()
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 297, in _bootstrap
    self.run()
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 99, in run
    self._target(*self._args, **self._kwargs)
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 99, in run
    self._target(*self._args, **self._kwargs)
  File "/home/ubuntu/_uti_basic.py", line 62, in fun
    q_out.put((i, f(x)))
  File "/home/ubuntu/_uti_basic.py", line 62, in fun
    q_out.put((i, f(x)))
  File "/home/ubuntu/main.py", line 85, in getOneRegionValue
    spatial = False)
  File "/home/ubuntu/main.py", line 85, in getOneRegionValue
    spatial = False)
  File "/home/ubuntu/main.py", line 236, in getWeight
    epsilon = epsilon, spatial = spatial, mean_field = mean_field)
  File "/home/ubuntu/main.py", line 236, in getWeight
    epsilon = epsilon, spatial = spatial, mean_field = mean_field)
  File "/home/ubuntu/weight.py", line 297, in train
    self.policy_ratio2: policy_ratio2
  File "/home/ubuntu/weight.py", line 297, in train
    self.policy_ratio2: policy_ratio2
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 950, in run
    run_metadata_ptr)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 950, in run
    run_metadata_ptr)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1173, in _run
    feed_dict_tensor, options, run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1173, in _run
    feed_dict_tensor, options, run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1350, in _do_run
    run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1350, in _do_run
    run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1356, in _do_call
    return fn(*args)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1356, in _do_call
    return fn(*args)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1341, in _run_fn
    options, feed_dict, fetch_list, target_list, run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1341, in _run_fn
    options, feed_dict, fetch_list, target_list, run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1429, in _call_tf_sessionrun
    run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1429, in _call_tf_sessionrun
    run_metadata)
KeyboardInterrupt
KeyboardInterrupt
Traceback (most recent call last):
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 297, in _bootstrap
    self.run()
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/process.py", line 99, in run
    self._target(*self._args, **self._kwargs)
  File "/home/ubuntu/_uti_basic.py", line 62, in fun
    q_out.put((i, f(x)))
  File "/home/ubuntu/main.py", line 85, in getOneRegionValue
    spatial = False)
  File "/home/ubuntu/main.py", line 236, in getWeight
    epsilon = epsilon, spatial = spatial, mean_field = mean_field)
  File "/home/ubuntu/weight.py", line 297, in train
    self.policy_ratio2: policy_ratio2
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 950, in run
    run_metadata_ptr)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1173, in _run
    feed_dict_tensor, options, run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1350, in _do_run
    run_metadata)
Traceback (most recent call last):
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1356, in _do_call
    return fn(*args)
```

```
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1341, in _run_fn
    options, feed_dict, fetch_list, target_list, run_metadata)
  File "/home/ubuntu/.local/lib/python3.7/site-packages/tensorflow/python/client/session.py", line 1429, in _call_tf_sessionrun
    run_metadata)
Process Process-193:
ubuntu@ip-172-31-15-241:~$ python EC2.py
21:52, 03/29; num of cores:16

Basic setting:[sd_O, sd_D, sd_R, sd_u_O, w_O, w_A, lam] = [4, 4, 3, 0.4, 1, 1, 0.0001]


---------------------------------------
[pattern_seed, T, sd_R] = [0, 672, 3]

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 = 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
---------------------------------------
Value of Behaviour policy:9.452
O_threshold = 12
MC for this TARGET:[10.145, 0.023]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.01, -0.03, 0.0]][[0.05, 0.03, 0.05]][[-10.14, -10.14, -10.14]][[-0.03, -0.69]]
std:[[0.06, 0.05, 0.1]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.08, 0.0]]
MSE:[[0.06, 0.06, 0.1]][[0.05, 0.03, 0.05]][[10.14, 10.14, 10.14]][[0.09, 0.69]]
MSE(-DR):[[0.0, 0.0, 0.04]][[-0.01, -0.03, -0.01]][[10.08, 10.08, 10.08]][[0.03, 0.63]]
==============
O_threshold = 9
MC for this TARGET:[9.859, 0.025]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.32, 0.3, 0.33]][[0.39, 0.37, 0.38]][[-9.86, -9.86, -9.86]][[0.31, -0.41]]
std:[[0.0, 0.02, 0.04]][[0.02, 0.02, 0.02]][[0.0, 0.0, 0.0]][[0.02, 0.0]]
MSE:[[0.32, 0.3, 0.33]][[0.39, 0.37, 0.38]][[9.86, 9.86, 9.86]][[0.31, 0.41]]
MSE(-DR):[[0.0, -0.02, 0.01]][[0.07, 0.05, 0.06]][[9.54, 9.54, 9.54]][[-0.01, 0.09]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
MC-based ATE = -0.29
```

```
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.31, 0.32, 0.33]][[0.35, 0.35, 0.34]][[0.29, 0.29, 0.29]][0.34]
std:[[0.07, 0.07, 0.06]][[0.0, 0.01, 0.0]][[0.0, 0.0, 0.0]][0.07]
MSE:[[0.32, 0.33, 0.34]][[0.35, 0.35, 0.34]][[0.29, 0.29, 0.29]][0.35]
MSE(-DR):[[0.0, 0.01, 0.02]][[0.03, 0.03, 0.02]][[-0.03, -0.03, -0.03]][0.03]
***** BETTER THAN [IS, DR_NO_MARL] *****
==============
O_threshold = 15
MC for this TARGET:[10.218, 0.024]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.44, -0.46, -0.42]][[-0.54, -0.56, -0.55]][[-10.22, -10.22, -10.22]][[-0.44, -0.77]]
std:[[0.0, 0.0, 0.05]][[0.02, 0.02, 0.01]][[0.0, 0.0, 0.0]][[0.04, 0.0]]
MSE:[[0.44, 0.46, 0.42]][[0.54, 0.56, 0.55]][[10.22, 10.22, 10.22]][[0.44, 0.77]]
MSE(-DR):[[0.0, 0.02, -0.02]][[0.1, 0.12, 0.11]][[9.78, 9.78, 9.78]][[0.0, 0.33]]
better than DR_NO_MARL
MC-based ATE = 0.07
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.45, -0.43, -0.43]][[-0.59, -0.58, -0.59]][[-0.07, -0.07, -0.07]][-0.41]
std:[[0.06, 0.05, 0.05]][[0.01, 0.01, 0.0]][[0.0, 0.0, 0.0]][0.04]
MSE:[[0.45, 0.43, 0.43]][[0.59, 0.58, 0.59]][[0.07, 0.07, 0.07]][0.41]
MSE(-DR):[[0.0, -0.02, -0.02]][[0.14, 0.13, 0.14]][[-0.38, -0.38, -0.38]][-0.04]
better than DR_NO_MARL
==============
time spent until now: 2.5 mins


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

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 = 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
---------------------------------------
Value of Behaviour policy:7.246
O_threshold = 12
MC for this TARGET:[7.834, 0.023]
```

```
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.13, -0.14, -0.15]][[-0.29, -0.31, -0.29]][[-7.83, -7.83, -7.83]][[-0.15, -0.59]]
std:[[0.04, 0.05, 0.02]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.03, 0.0]]
MSE:[[0.14, 0.15, 0.15]][[0.29, 0.31, 0.29]][[7.83, 7.83, 7.83]][[0.15, 0.59]]
MSE(-DR):[[0.0, 0.01, 0.01]][[0.15, 0.17, 0.15]][[7.69, 7.69, 7.69]][[0.01, 0.45]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
==============
O_threshold = 9
MC for this TARGET:[7.683, 0.023]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.15, 0.14, 0.16]][[0.12, 0.11, 0.13]][[-7.68, -7.68, -7.68]][[0.14, -0.44]]
std:[[0.08, 0.07, 0.09]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.09, 0.0]]
MSE:[[0.17, 0.16, 0.18]][[0.12, 0.11, 0.13]][[7.68, 7.68, 7.68]][[0.17, 0.44]]
MSE(-DR):[[0.0, -0.01, 0.01]][[-0.05, -0.06, -0.04]][[7.51, 7.51, 7.51]][[0.0, 0.27]]
MC-based ATE = -0.15
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.29, 0.28, 0.31]][[0.41, 0.41, 0.42]][[0.15, 0.15, 0.15]][0.29]
std:[[0.04, 0.02, 0.08]][[0.0, 0.0, 0.01]][[0.0, 0.0, 0.0]][0.06]
MSE:[[0.29, 0.28, 0.32]][[0.41, 0.41, 0.42]][[0.15, 0.15, 0.15]][0.3]
MSE(-DR):[[0.0, -0.01, 0.03]][[0.12, 0.12, 0.13]][[-0.14, -0.14, -0.14]][0.01]
***** BETTER THAN [IS, DR_NO_MARL] *****
==============
O_threshold = 15
MC for this TARGET:[7.787, 0.023]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.38, -0.38, -0.3]][[-0.5, -0.51, -0.5]][[-7.79, -7.79, -7.79]][[-0.3, -0.54]]
std:[[0.01, 0.01, 0.02]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.01, 0.0]]
MSE:[[0.38, 0.38, 0.3]][[0.5, 0.51, 0.5]][[7.79, 7.79, 7.79]][[0.3, 0.54]]
MSE(-DR):[[0.0, 0.0, -0.08]][[0.12, 0.13, 0.12]][[7.41, 7.41, 7.41]][[-0.08, 0.16]]
better than DR_NO_MARL
MC-based ATE = -0.05
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.24, -0.24, -0.15]][[-0.21, -0.2, -0.21]][[0.05, 0.05, 0.05]][-0.15]
std:[[0.03, 0.04, 0.03]][[0.02, 0.01, 0.01]][[0.0, 0.0, 0.0]][0.04]
MSE:[[0.24, 0.24, 0.15]][[0.21, 0.2, 0.21]][[0.05, 0.05, 0.05]][0.16]
MSE(-DR):[[0.0, 0.0, -0.09]][[-0.03, -0.04, -0.03]][[-0.19, -0.19, -0.19]][-0.08]
==============
time spent until now: 4.9 mins


---------------------------------------
[pattern_seed, T, sd_R] = [2, 672, 3]

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 = 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 MSE:[[0.32, 0.33, 0.34]][[0.35, 0.35, 0.34]][[0.29, 0.29, 0.29]][0.35]
2 3
----------------------------------------
Value of Behaviour policy:6.7
O_threshold = 12
MC for this TARGET:[7.138, 0.023]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.42, -0.4, -0.38]][[-0.48, -0.49, -0.48]][[-7.14, -7.14, -7.14]][[-0.37, -0.44]]
std:[[0.08, 0.07, 0.05]][[0.05, 0.05, 0.04]][[0.0, 0.0, 0.0]][[0.04, 0.0]]
MSE:[[0.43, 0.41, 0.38]][[0.48, 0.49, 0.48]][[7.14, 7.14, 7.14]][[0.37, 0.44]]
MSE(-DR):[[0.0, -0.02, -0.05]][[0.05, 0.06, 0.05]][[6.71, 6.71, 6.71]][[-0.06, 0.01]]
better than DR_NO_MARL
==============
O_threshold = 9
MC for this TARGET:[7.178, 0.023]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.02, -0.02, -0.08]][[-0.09, -0.11, -0.09]][[-7.18, -7.18, -7.18]][[-0.08, -0.48]]
std:[[0.11, 0.12, 0.05]][[0.05, 0.05, 0.04]][[0.0, 0.0, 0.0]][[0.06, 0.0]]
MSE:[[0.11, 0.12, 0.09]][[0.1, 0.12, 0.1]][[7.18, 7.18, 7.18]][[0.1, 0.48]]
MSE(-DR):[[0.0, 0.01, -0.02]][[-0.01, 0.01, -0.01]][[7.07, 7.07, 7.07]][[-0.01, 0.37]]
MC-based ATE = 0.04
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.4, 0.38, 0.3]][[0.39, 0.38, 0.39]][[-0.04, -0.04, -0.04]][0.28]
std:[[0.03, 0.05, 0.0]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.02]
MSE:[[0.4, 0.38, 0.3]][[0.39, 0.38, 0.39]][[0.04, 0.04, 0.04]][0.28]
MSE(-DR):[[0.0, -0.02, -0.1]][[-0.01, -0.02, -0.01]][[-0.36, -0.36, -0.36]][-0.12]
==============
O_threshold = 15
MC for this TARGET:[6.995, 0.023]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.55, -0.54, -0.56]][[-0.66, -0.66, -0.66]][[-7.0, -7.0, -7.0]][[-0.55, -0.29]]
std:[[0.05, 0.04, 0.01]][[0.06, 0.05, 0.04]][[0.0, 0.0, 0.0]][[0.0, 0.0]]
MSE:[[0.55, 0.54, 0.56]][[0.66, 0.66, 0.66]][[7.0, 7.0, 7.0]][[0.55, 0.29]]
MSE(-DR):[[0.0, -0.01, 0.01]][[0.11, 0.11, 0.11]][[6.45, 6.45, 6.45]][[0.0, -0.26]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
MC-based ATE = -0.14
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.13, -0.14, -0.19]][[-0.18, -0.18, -0.18]][[0.14, 0.14, 0.14]][-0.19]
std:[[0.03, 0.03, 0.04]][[0.01, 0.01, 0.0]][[0.0, 0.0, 0.0]][0.04]
MSE:[[0.13, 0.14, 0.19]][[0.18, 0.18, 0.18]][[0.14, 0.14, 0.14]][0.19]
MSE(-DR):[[0.0, 0.01, 0.06]][[0.05, 0.05, 0.05]][[0.01, 0.01, 0.01]][0.06]
***** BETTER THAN [IS, DR_NO_MARL] *****
==============
time spent until now: 7.4 mins


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

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 = 9
target policy:

1 1 1 0 1
```

```
1 1 0 1 1

0 1 1 1 1

1 0 0 1 0

0 1 1 1 0

number of reward locations:  17
O_threshold = 15
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 ^CProcess Process-512:
Process Process-505:
Traceback (most recent call last):
  File "EC2.py", line 68, in <module>
Process Process-509:
Process Process-497:
Process Process-500:
Process Process-501:
    file = file, print_flag_target = False
  File "/home/ubuntu/simu_funs.py", line 62, in simu
    value_reps = rep_seeds(once, OPE_rep_times)
  File "/home/ubuntu/_uti_basic.py", line 119, in rep_seeds
    return list(map(fun, range(rep_times)))
  File "/home/ubuntu/simu_funs.py", line 58, in once
    inner_parallel = inner_parallel)
  File "/home/ubuntu/simu_funs.py", line 191, in simu_once
    inner_parallel = inner_parallel)
  File "/home/ubuntu/main.py", line 130, in V_DR
    r = arr(parmap(getOneRegionValue, range(N), n_cores))
  File "/home/ubuntu/_uti_basic.py", line 75, in parmap
    [q_in.put((None, None)) for _ in range(nprocs)]
  File "/home/ubuntu/_uti_basic.py", line 75, in <listcomp>
Process Process-498:
    [q_in.put((None, None)) for _ in range(nprocs)]
  File "/home/ubuntu/anaconda3/lib/python3.7/multiprocessing/queues.py", line 82, in put
    if not self._sem.acquire(block, timeout):
KeyboardInterrupt
ubuntu@ip-172-31-15-241:~$ python EC2.py
22:01, 03/29; num of cores:16

Basic setting:[sd_O, sd_D, sd_R, sd_u_O, w_O, w_A, lam] = [2, 2, 2, 0.4, 1, 1, 0.0001]


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

max(u_O) =  27.0
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

target policy:

1 1 1 1

1 0 1 0

0 1 0 1

1 0 1 1

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

1 1 1 1
```

```
1 0 1 1

1 1 1 1

1 1 1 1

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

1 0 1 1

1 0 1 0

0 0 0 1

0 0 0 0

number of reward locations:  6
1 2 3 1 2 3
----------------------------------------
Value of Behaviour policy:10.388
O_threshold = 12
MC for this TARGET:[11.131, 0.019]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.08, 0.06, 0.0]][[0.1, 0.07, 0.08]][[-11.13, -11.13, -11.13]][[-0.01, -0.74]]
std:[[0.09, 0.08, 0.06]][[0.01, 0.02, 0.03]][[0.0, 0.0, 0.0]][[0.05, 0.01]]
MSE:[[0.12, 0.1, 0.06]][[0.1, 0.07, 0.09]][[11.13, 11.13, 11.13]][[0.05, 0.74]]
MSE(-DR):[[0.0, -0.02, -0.06]][[-0.02, -0.05, -0.03]][[11.01, 11.01, 11.01]][[-0.07, 0.62]]
==============
O_threshold = 9
MC for this TARGET:[10.703, 0.02]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.35, 0.34, 0.27]][[0.47, 0.44, 0.45]][[-10.7, -10.7, -10.7]][[0.25, -0.31]]
std:[[0.08, 0.08, 0.06]][[0.03, 0.03, 0.02]][[0.0, 0.0, 0.0]][[0.06, 0.01]]
MSE:[[0.36, 0.35, 0.28]][[0.47, 0.44, 0.45]][[10.7, 10.7, 10.7]][[0.26, 0.31]]
MSE(-DR):[[0.0, -0.01, -0.08]][[0.11, 0.08, 0.09]][[10.34, 10.34, 10.34]][[-0.1, -0.05]]
better than DR_NO_MARL
MC-based ATE = -0.43
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.27, 0.27, 0.26]][[0.37, 0.37, 0.37]][[0.43, 0.43, 0.43]][0.26]
std:[[0.01, 0.0, 0.0]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][0.0]
MSE:[[0.27, 0.27, 0.26]][[0.37, 0.37, 0.37]][[0.43, 0.43, 0.43]][0.26]
MSE(-DR):[[0.0, 0.0, -0.01]][[0.1, 0.1, 0.1]][[0.16, 0.16, 0.16]][-0.01]
better than DR_NO_MARL
==============
O_threshold = 15
MC for this TARGET:[11.228, 0.02]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.26, -0.28, -0.35]][[-0.46, -0.48, -0.48]][[-11.23, -11.23, -11.23]][[-0.37, -0.84]]
std:[[0.01, 0.01, 0.02]][[0.02, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.01, 0.01]]
MSE:[[0.26, 0.28, 0.35]][[0.46, 0.48, 0.48]][[11.23, 11.23, 11.23]][[0.37, 0.84]]
MSE(-DR):[[0.0, 0.02, 0.09]][[0.2, 0.22, 0.22]][[10.97, 10.97, 10.97]][[0.11, 0.58]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
MC-based ATE = 0.1
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.34, -0.35, -0.35]][[-0.56, -0.55, -0.56]][[-0.1, -0.1, -0.1]][-0.36]
std:[[0.07, 0.07, 0.05]][[0.03, 0.03, 0.04]][[0.0, 0.0, 0.0]][0.04]
MSE:[[0.35, 0.36, 0.35]][[0.56, 0.55, 0.56]][[0.1, 0.1, 0.1]][0.36]
MSE(-DR):[[0.0, 0.01, 0.0]][[0.21, 0.2, 0.21]][[-0.25, -0.25, -0.25]][0.01]
***** BETTER THAN [IS, DR_NO_MARL] *****
==============
time spent until now: 1.5 mins


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

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

target policy:

1 0 0 0

1 0 1 0

1 0 1 0

0 0 1 0
```

```
number of reward locations:  6
O_threshold = 9
target policy:

1 0 0 0

1 0 1 0

1 1 1 0

1 1 1 0

number of reward locations:  9
O_threshold = 15
target policy:

1 0 0 0

1 0 1 0

0 0 1 0

0 0 1 0

number of reward locations:  5
1 2 3 1 2 3
----------------------------------------
Value of Behaviour policy:7.11
O_threshold = 12
MC for this TARGET:[7.811, 0.02]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.48, -0.48, -0.47]][[-0.4, -0.42, -0.41]][[-7.81, -7.81, -7.81]][[-0.48, -0.7]]
std:[[0.01, 0.0, 0.01]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.01, 0.01]]
MSE:[[0.48, 0.48, 0.47]][[0.4, 0.42, 0.41]][[7.81, 7.81, 7.81]][[0.48, 0.7]]
MSE(-DR):[[0.0, 0.0, -0.01]][[-0.08, -0.06, -0.07]][[7.33, 7.33, 7.33]][[0.0, 0.22]]
==============
O_threshold = 9
MC for this TARGET:[7.691, 0.021]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.13, -0.13, -0.17]][[-0.04, -0.07, -0.06]][[-7.69, -7.69, -7.69]][[-0.18, -0.58]]
std:[[0.04, 0.03, 0.02]][[0.0, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.02, 0.01]]
MSE:[[0.14, 0.13, 0.17]][[0.04, 0.07, 0.06]][[7.69, 7.69, 7.69]][[0.18, 0.58]]
MSE(-DR):[[0.0, -0.01, 0.03]][[-0.1, -0.07, -0.08]][[7.55, 7.55, 7.55]][[0.04, 0.44]]
MC-based ATE = -0.12
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.35, 0.35, 0.3]][[0.35, 0.35, 0.35]][[0.12, 0.12, 0.12]][0.31]
std:[[0.03, 0.03, 0.01]][[0.01, 0.01, 0.0]][[0.0, 0.0, 0.0]][0.01]
MSE:[[0.35, 0.35, 0.3]][[0.35, 0.35, 0.35]][[0.12, 0.12, 0.12]][0.31]
MSE(-DR):[[0.0, 0.0, -0.05]][[0.0, 0.0, 0.0]][[-0.23, -0.23, -0.23]][-0.04]
better than DR_NO_MARL
==============
O_threshold = 15
MC for this TARGET:[7.77, 0.02]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.53, -0.53, -0.56]][[-0.5, -0.52, -0.5]][[-7.77, -7.77, -7.77]][[-0.56, -0.66]]
std:[[0.02, 0.02, 0.01]][[0.0, 0.0, 0.01]][[0.0, 0.0, 0.0]][[0.01, 0.01]]
MSE:[[0.53, 0.53, 0.56]][[0.5, 0.52, 0.5]][[7.77, 7.77, 7.77]][[0.56, 0.66]]
MSE(-DR):[[0.0, 0.0, 0.03]][[-0.03, -0.01, -0.03]][[7.24, 7.24, 7.24]][[0.03, 0.13]]
MC-based ATE = -0.04
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.05, -0.05, -0.08]][[-0.1, -0.1, -0.1]][[0.04, 0.04, 0.04]][-0.08]
std:[[0.03, 0.03, 0.02]][[0.01, 0.0, 0.01]][[0.0, 0.0, 0.0]][0.02]
MSE:[[0.06, 0.06, 0.08]][[0.1, 0.1, 0.1]][[0.04, 0.04, 0.04]][0.08]
MSE(-DR):[[0.0, 0.0, 0.02]][[0.04, 0.04, 0.04]][[-0.02, -0.02, -0.02]][0.02]
***** BETTER THAN [IS, DR_NO_MARL] *****
==============
time spent until now: 3.1 mins


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

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

target policy:

0 0 0 1
```

```
0 0 1 0

0 0 1 1

0 0 1 0

number of reward locations:  5
O_threshold = 9
target policy:

1 1 0 1

0 0 1 0

0 0 1 1

1 0 1 0

number of reward locations:  8
O_threshold = 15
target policy:

0 0 0 1

0 0 0 0

0 0 0 1

0 0 0 0

number of reward locations:  2
1 2 3 1 2 3
----------------------------------------
Value of Behaviour policy:6.558
O_threshold = 12
MC for this TARGET:[7.058, 0.02]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.24, -0.24, -0.25]][[-0.37, -0.37, -0.38]][[-7.06, -7.06, -7.06]][[-0.26, -0.5]]
std:[[0.03, 0.02, 0.04]][[0.0, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.03, 0.01]]
MSE:[[0.24, 0.24, 0.25]][[0.37, 0.37, 0.38]][[7.06, 7.06, 7.06]][[0.26, 0.5]]
MSE(-DR):[[0.0, 0.0, 0.01]][[0.13, 0.13, 0.14]][[6.82, 6.82, 6.82]][[0.02, 0.26]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
==============
O_threshold = 9
MC for this TARGET:[7.209, 0.02]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.19, -0.21, -0.22]][[-0.22, -0.23, -0.23]][[-7.21, -7.21, -7.21]][[-0.23, -0.65]]
std:[[0.06, 0.06, 0.04]][[0.02, 0.03, 0.02]][[0.0, 0.0, 0.0]][[0.04, 0.01]]
MSE:[[0.2, 0.22, 0.22]][[0.22, 0.23, 0.23]][[7.21, 7.21, 7.21]][[0.23, 0.65]]
MSE(-DR):[[0.0, 0.02, 0.02]][[0.02, 0.03, 0.03]][[7.01, 7.01, 7.01]][[0.03, 0.45]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
MC-based ATE = 0.15
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.05, 0.03, 0.04]][[0.15, 0.14, 0.15]][[-0.15, -0.15, -0.15]][0.02]
std:[[0.04, 0.05, 0.0]][[0.02, 0.02, 0.02]][[0.0, 0.0, 0.0]][0.01]
MSE:[[0.06, 0.06, 0.04]][[0.15, 0.14, 0.15]][[0.15, 0.15, 0.15]][0.02]
MSE(-DR):[[0.0, 0.0, -0.02]][[0.09, 0.08, 0.09]][[0.09, 0.09, 0.09]][-0.04]
better than DR_NO_MARL
==============
O_threshold = 15
MC for this TARGET:[6.787, 0.02]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.25, -0.24, -0.26]][[-0.55, -0.54, -0.55]][[-6.79, -6.79, -6.79]][[-0.25, -0.23]]
std:[[0.01, 0.02, 0.0]][[0.01, 0.02, 0.02]][[0.0, 0.0, 0.0]][[0.0, 0.01]]
MSE:[[0.25, 0.24, 0.26]][[0.55, 0.54, 0.55]][[6.79, 6.79, 6.79]][[0.25, 0.23]]
MSE(-DR):[[0.0, -0.01, 0.01]][[0.3, 0.29, 0.3]][[6.54, 6.54, 6.54]][[0.0, -0.02]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
MC-based ATE = -0.27
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.01, -0.0, -0.0]][[-0.18, -0.17, -0.17]][[0.27, 0.27, 0.27]][0.0]
std:[[0.04, 0.03, 0.04]][[0.02, 0.02, 0.01]][[0.0, 0.0, 0.0]][0.03]
MSE:[[0.04, 0.03, 0.04]][[0.18, 0.17, 0.17]][[0.27, 0.27, 0.27]][0.03]
MSE(-DR):[[0.0, -0.01, 0.0]][[0.14, 0.13, 0.13]][[0.23, 0.23, 0.23]][-0.01]
***** BETTER THAN [IS, DR_NO_MARL] *****
==============
time spent until now: 4.6 mins


----------------------------------------
[pattern_seed, T, sd_R] = [3, 672, 2]

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

target policy:

1 1 0 0

0 0 0 0

0 0 0 1

1 1 0 0

number of reward locations:  5
O_threshold = 9
target policy:

1 1 1 0

1 1 1 0

1 1 0 1

1 1 1 1

number of reward locations:  13
O_threshold = 15
target policy:

1 0 0 0

0 0 0 0

0 0 0 1

1 1 0 0

number of reward locations:  4
1 2 3 1 2 3
----------------------------------------
Value of Behaviour policy:7.76
O_threshold = 12
MC for this TARGET:[8.334, 0.018]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.51, -0.51, -0.51]][[-0.53, -0.54, -0.53]][[-8.33, -8.33, -8.33]][[-0.52, -0.57]]
std:[[0.07, 0.07, 0.06]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.07, 0.01]]
MSE:[[0.51, 0.51, 0.51]][[0.53, 0.54, 0.53]][[8.33, 8.33, 8.33]][[0.52, 0.57]]
MSE(-DR):[[0.0, 0.0, 0.0]][[0.02, 0.03, 0.02]][[7.82, 7.82, 7.82]][[0.01, 0.06]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
==============
O_threshold = 9
MC for this TARGET:[8.158, 0.019]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.12, 0.11, 0.12]][[0.26, 0.23, 0.25]][[-8.16, -8.16, -8.16]][[0.11, -0.4]]
std:[[0.06, 0.06, 0.02]][[0.0, 0.0, 0.01]][[0.0, 0.0, 0.0]][[0.02, 0.01]]
MSE:[[0.13, 0.13, 0.12]][[0.26, 0.23, 0.25]][[8.16, 8.16, 8.16]][[0.11, 0.4]]
MSE(-DR):[[0.0, 0.0, -0.01]][[0.13, 0.1, 0.12]][[8.03, 8.03, 8.03]][[-0.02, 0.27]]
better than DR_NO_MARL
MC-based ATE = -0.18
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.64, 0.63, 0.63]][[0.79, 0.78, 0.78]][[0.18, 0.18, 0.18]][0.63]
std:[[0.13, 0.13, 0.08]][[0.0, 0.01, 0.03]][[0.0, 0.0, 0.0]][0.09]
MSE:[[0.65, 0.64, 0.64]][[0.79, 0.78, 0.78]][[0.18, 0.18, 0.18]][0.64]
MSE(-DR):[[0.0, -0.01, -0.01]][[0.14, 0.13, 0.13]][[-0.47, -0.47, -0.47]][-0.01]
better than DR_NO_MARL
==============
O_threshold = 15
MC for this TARGET:[8.272, 0.019]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.57, -0.56, -0.52]][[-0.6, -0.62, -0.61]][[-8.27, -8.27, -8.27]][[-0.52, -0.51]]
std:[[0.07, 0.06, 0.07]][[0.01, 0.01, 0.02]][[0.0, 0.0, 0.0]][[0.07, 0.01]]
MSE:[[0.57, 0.56, 0.52]][[0.6, 0.62, 0.61]][[8.27, 8.27, 8.27]][[0.52, 0.51]]
MSE(-DR):[[0.0, -0.01, -0.05]][[0.03, 0.05, 0.04]][[7.7, 7.7, 7.7]][[-0.05, -0.06]]
better than DR_NO_MARL
MC-based ATE = -0.06
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.05, -0.05, -0.0]][[-0.08, -0.07, -0.08]][[0.06, 0.06, 0.06]][-0.0]
std:[[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.0]
MSE:[[0.05, 0.05, 0.0]][[0.08, 0.07, 0.08]][[0.06, 0.06, 0.06]][0.0]
MSE(-DR):[[0.0, 0.0, -0.05]][[0.03, 0.02, 0.03]][[0.01, 0.01, 0.01]][-0.05]
better than DR_NO_MARL
==============
time spent until now: 6.2 mins


----------------------------------------
[pattern_seed, T, sd_R] = [4, 672, 2]
```

```
max(u_O) =  14.5
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

target policy:

0 1 0 1

0 0 0 1

1 0 1 0

1 1 0 1

number of reward locations:  8
O_threshold = 9
target policy:

1 1 0 1

1 0 0 1

1 0 1 1

1 1 0 1

number of reward locations:  11
O_threshold = 15
target policy:

0 0 0 0

0 0 0 0

0 0 0 0

0 0 0 0

number of reward locations:  0
1
```