```
Last login: Sun Mar 29 21:15:09 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
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 disabled due to load higher than 16.0

 * 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 01:15:28 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
22:09, 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) =  27.3
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 6.2 11.2 10.2 20.4 19.8

11.7 12.8 7.7 5.0 9.6 11.7

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 0 0 0 1 1

0 1 0 0 0 0

number of reward locations:  19
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 0 1 1 1 1

1 1 0 0 1 1

number of reward locations:  29
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 0 0 0 1 1

0 0 0 0 0 0

number of reward locations:  11
1 2 3 1 2 3
--------------------------------------
Value of Behaviour policy:8.823
O_threshold = 12
MC for this TARGET:[9.502, 0.014]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.09, -0.1, -0.12]][[-0.04, -0.06, -0.06]][[-9.5, -9.5, -9.5]][[-0.13, -0.68]]
std:[[0.01, 0.01, 0.05]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.05, 0.0]]
MSE:[[0.09, 0.1, 0.13]][[0.04, 0.06, 0.06]][[9.5, 9.5, 9.5]][[0.14, 0.68]]
MSE(-DR):[[0.0, 0.01, 0.04]][[-0.05, -0.03, -0.03]][[9.41, 9.41, 9.41]][[0.05, 0.59]]
==============
O_threshold = 9
MC for this TARGET:[9.28, 0.014]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.31, 0.3, 0.24]][[0.38, 0.36, 0.37]][[-9.28, -9.28, -9.28]][[0.22, -0.46]]
std:[[0.04, 0.04, 0.04]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.04, 0.0]]
MSE:[[0.31, 0.3, 0.24]][[0.38, 0.36, 0.37]][[9.28, 9.28, 9.28]][[0.22, 0.46]]
MSE(-DR):[[0.0, -0.01, -0.07]][[0.07, 0.05, 0.06]][[8.97, 8.97, 8.97]][[-0.09, 0.15]]
better than DR_NO_MARL
MC-based ATE = -0.22
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.4, 0.41, 0.35]][[0.42, 0.42, 0.43]][[0.22, 0.22, 0.22]][0.35]
std:[[0.03, 0.03, 0.0]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][0.01]
MSE:[[0.4, 0.41, 0.35]][[0.42, 0.42, 0.43]][[0.22, 0.22, 0.22]][0.35]
MSE(-DR):[[0.0, 0.01, -0.05]][[0.02, 0.02, 0.03]][[-0.18, -0.18, -0.18]][-0.05]
better than DR_NO_MARL
==============
O_threshold = 15
MC for this TARGET:[9.439, 0.014]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.44, -0.45, -0.46]][[-0.53, -0.54, -0.54]][[-9.44, -9.44, -9.44]][[-0.47, -0.62]]
std:[[0.05, 0.04, 0.01]][[0.02, 0.02, 0.02]][[0.0, 0.0, 0.0]][[0.01, 0.0]]
MSE:[[0.44, 0.45, 0.46]][[0.53, 0.54, 0.54]][[9.44, 9.44, 9.44]][[0.47, 0.62]]
MSE(-DR):[[0.0, 0.01, 0.02]][[0.09, 0.1, 0.1]][[9.0, 9.0, 9.0]][[0.03, 0.18]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
MC-based ATE = -0.06
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.35, -0.35, -0.34]][[-0.49, -0.48, -0.48]][[0.06, 0.06, 0.06]][-0.34]
std:[[0.05, 0.05, 0.04]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][0.03]
MSE:[[0.35, 0.35, 0.34]][[0.49, 0.48, 0.48]][[0.06, 0.06, 0.06]][0.34]
MSE(-DR):[[0.0, 0.0, -0.01]][[0.14, 0.13, 0.13]][[-0.29, -0.29, -0.29]][-0.01]
better than DR_NO_MARL
==============
time spent until now: 3.6 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 10.3 7.8

11.2 13.9 7.1 17.4 15.8 13.5

15.8 8.4 10.5 7.6 9.9 13.6

8.4 9.4 8.4 7.9 8.4 11.0

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 0 0 0 0 1

0 0 0 0 0 0

number of reward locations:  12
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 0 1 0 1 1

0 1 0 0 0 1

number of reward locations:  21
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 0 0 0 0 0

0 0 0 0 0 0

number of reward locations:  8
1 2 3 1 2 3
--------------------------------------
Value of Behaviour policy:7.184
O_threshold = 12
MC for this TARGET:[7.718, 0.014]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.35, -0.37, -0.35]][[-0.43, -0.44, -0.43]][[-7.72, -7.72, -7.72]][[-0.37, -0.53]]
std:[[0.07, 0.07, 0.03]][[0.01, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.03, 0.01]]
MSE:[[0.36, 0.38, 0.35]][[0.43, 0.44, 0.43]][[7.72, 7.72, 7.72]][[0.37, 0.53]]
MSE(-DR):[[0.0, 0.02, -0.01]][[0.07, 0.08, 0.07]][[7.36, 7.36, 7.36]][[0.01, 0.17]]
better than DR_NO_MARL
==============
O_threshold = 9
MC for this TARGET:[7.669, 0.014]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.05, -0.06, -0.05]][[0.0, -0.02, -0.0]][[-7.67, -7.67, -7.67]][[-0.07, -0.49]]
std:[[0.02, 0.03, 0.03]][[0.02, 0.02, 0.01]][[0.0, 0.0, 0.0]][[0.03, 0.01]]
MSE:[[0.05, 0.07, 0.06]][[0.02, 0.03, 0.01]][[7.67, 7.67, 7.67]][[0.08, 0.49]]
MSE(-DR):[[0.0, 0.02, 0.01]][[-0.03, -0.02, -0.04]][[7.62, 7.62, 7.62]][[0.03, 0.44]]
MC-based ATE = -0.05
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.3, 0.31, 0.3]][[0.43, 0.42, 0.43]][[0.05, 0.05, 0.05]][0.3]
std:[[0.1, 0.1, 0.06]][[0.02, 0.02, 0.01]][[0.0, 0.0, 0.0]][0.06]
MSE:[[0.32, 0.33, 0.31]][[0.43, 0.42, 0.43]][[0.05, 0.05, 0.05]][0.31]
MSE(-DR):[[0.0, 0.01, -0.01]][[0.11, 0.1, 0.11]][[-0.27, -0.27, -0.27]][-0.01]
better than DR_NO_MARL
==============
O_threshold = 15
MC for this TARGET:[7.63, 0.014]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.48, -0.48, -0.46]][[-0.59, -0.6, -0.59]][[-7.63, -7.63, -7.63]][[-0.47, -0.45]]
std:[[0.06, 0.06, 0.02]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.02, 0.01]]
MSE:[[0.48, 0.48, 0.46]][[0.59, 0.6, 0.59]][[7.63, 7.63, 7.63]][[0.47, 0.45]]
MSE(-DR):[[0.0, 0.0, -0.02]][[0.11, 0.12, 0.11]][[7.15, 7.15, 7.15]][[-0.01, -0.03]]
better than DR_NO_MARL
MC-based ATE = -0.09
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.12, -0.11, -0.1]][[-0.16, -0.16, -0.16]][[0.09, 0.09, 0.09]][-0.1]
std:[[0.02, 0.02, 0.01]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][0.01]
MSE:[[0.12, 0.11, 0.1]][[0.16, 0.16, 0.16]][[0.09, 0.09, 0.09]][0.1]
MSE(-DR):[[0.0, -0.01, -0.02]][[0.04, 0.04, 0.04]][[-0.03, -0.03, -0.03]][-0.02]
better than DR_NO_MARL
==============
time spent until now: 7.2 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 10.9 17.6

8.2 11.1 7.8 10.4 12.2 7.4

9.6 10.0 8.5 6.9 6.2 10.4

9.9 26.9 4.2 11.5 12.8 19.0

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 0 0 0 0 0

0 1 0 0 1 1

number of reward locations:  10
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 1 0 0 0 1

1 1 0 1 1 1

number of reward locations:  21
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 0 0 0 0 0

0 1 0 0 0 1

number of reward locations:  5
1 2 3 1 2 3
--------------------------------------
Value of Behaviour policy:6.863
O_threshold = 12
MC for this TARGET:[7.324, 0.013]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.42, -0.43, -0.38]][[-0.49, -0.5, -0.49]][[-7.32, -7.32, -7.32]][[-0.39, -0.46]]
std:[[0.05, 0.06, 0.0]][[0.01, 0.0, 0.01]][[0.0, 0.0, 0.0]][[0.01, 0.0]]
MSE:[[0.42, 0.43, 0.38]][[0.49, 0.5, 0.49]][[7.32, 7.32, 7.32]][[0.39, 0.46]]
MSE(-DR):[[0.0, 0.01, -0.04]][[0.07, 0.08, 0.07]][[6.9, 6.9, 6.9]][[-0.03, 0.04]]
better than DR_NO_MARL
==============
O_threshold = 9
MC for this TARGET:[7.434, 0.013]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.06, -0.08, -0.07]][[-0.03, -0.05, -0.04]][[-7.43, -7.43, -7.43]][[-0.09, -0.57]]
std:[[0.04, 0.04, 0.01]][[0.02, 0.02, 0.01]][[0.0, 0.0, 0.0]][[0.01, 0.0]]
MSE:[[0.07, 0.09, 0.07]][[0.04, 0.05, 0.04]][[7.43, 7.43, 7.43]][[0.09, 0.57]]
MSE(-DR):[[0.0, 0.02, 0.0]][[-0.03, -0.02, -0.03]][[7.36, 7.36, 7.36]][[0.02, 0.5]]
MC-based ATE = 0.11
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.36, 0.35, 0.31]][[0.46, 0.45, 0.46]][[-0.11, -0.11, -0.11]][0.3]
std:[[0.01, 0.02, 0.02]][[0.03, 0.02, 0.02]][[0.0, 0.0, 0.0]][0.01]
MSE:[[0.36, 0.35, 0.31]][[0.46, 0.45, 0.46]][[0.11, 0.11, 0.11]][0.3]
MSE(-DR):[[0.0, -0.01, -0.05]][[0.1, 0.09, 0.1]][[-0.25, -0.25, -0.25]][-0.06]
better than DR_NO_MARL
==============
O_threshold = 15
```

```
MC for this TARGET:[7.156, 0.014]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.49, -0.49, -0.48]][[-0.64, -0.64, -0.64]][[-7.16, -7.16, -7.16]][[-0.48, -0.29]]
std:[[0.02, 0.02, 0.01]][[0.0, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.01, 0.0]]
MSE:[[0.49, 0.49, 0.48]][[0.64, 0.64, 0.64]][[7.16, 7.16, 7.16]][[0.48, 0.29]]
MSE(-DR):[[0.0, 0.0, -0.01]][[0.15, 0.15, 0.15]][[6.67, 6.67, 6.67]][[-0.01, -0.2]]
better than DR_NO_MARL
MC-based ATE = -0.17
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.07, -0.06, -0.09]][[-0.15, -0.14, -0.14]][[0.17, 0.17, 0.17]][-0.08]
std:[[0.03, 0.04, 0.01]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.0]
MSE:[[0.08, 0.07, 0.09]][[0.15, 0.14, 0.14]][[0.17, 0.17, 0.17]][0.08]
MSE(-DR):[[0.0, -0.01, 0.01]][[0.07, 0.06, 0.06]][[0.09, 0.09, 0.09]][0.0]
***** BETTER THAN [IS, DR_NO_MARL] *****
==============
time spent until now: 10.7 mins


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

max(u_O) =  24.3
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 8.3 14.2 10.3 8.1 10.1

14.9 24.3 6.7 8.6 8.0 4.2

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 0 1 0 0 0

1 1 0 0 0 0

number of reward locations:  11
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 0 1 1 0 1

1 1 0 0 0 0

number of reward locations:  22
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 0 0 0 0 0

0 1 0 0 0 0

number of reward locations:  7
1 2 3 1 2 3
--------------------------------------
Value of Behaviour policy:7.026
O_threshold = 12
MC for this TARGET:[7.53, 0.014]
```

```
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.38, -0.37, -0.41]][[-0.47, -0.48, -0.46]][[-7.53, -7.53, -7.53]][[-0.4, -0.5]]
std:[[0.02, 0.02, 0.01]][[0.03, 0.03, 0.03]][[0.0, 0.0, 0.0]][[0.01, 0.01]]
MSE:[[0.38, 0.37, 0.41]][[0.47, 0.48, 0.46]][[7.53, 7.53, 7.53]][[0.4, 0.5]]
MSE(-DR):[[0.0, -0.01, 0.03]][[0.09, 0.1, 0.08]][[7.15, 7.15, 7.15]][[0.02, 0.12]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
==============
O_threshold = 9
MC for this TARGET:[7.544, 0.014]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.02, -0.03, -0.05]][[0.03, 0.01, 0.02]][[-7.54, -7.54, -7.54]][[-0.07, -0.52]]
std:[[0.01, 0.01, 0.01]][[0.02, 0.01, 0.02]][[0.0, 0.0, 0.0]][[0.01, 0.01]]
MSE:[[0.02, 0.03, 0.05]][[0.04, 0.01, 0.03]][[7.54, 7.54, 7.54]][[0.07, 0.52]]
MSE(-DR):[[0.0, 0.01, 0.03]][[0.02, -0.01, 0.01]][[7.52, 7.52, 7.52]][[0.05, 0.5]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
MC-based ATE = 0.01
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.36, 0.34, 0.35]][[0.5, 0.49, 0.49]][[-0.01, -0.01, -0.01]][0.33]
std:[[0.03, 0.03, 0.03]][[0.01, 0.02, 0.01]][[0.0, 0.0, 0.01]][0.02]
MSE:[[0.36, 0.34, 0.35]][[0.5, 0.49, 0.49]][[0.01, 0.01, 0.01]][0.33]
MSE(-DR):[[0.0, -0.02, -0.01]][[0.14, 0.13, 0.13]][[-0.35, -0.35, -0.35]][-0.03]
better than DR_NO_MARL
==============
O_threshold = 15
MC for this TARGET:[7.422, 0.014]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.44, -0.43, -0.43]][[-0.59, -0.59, -0.58]][[-7.42, -7.42, -7.42]][[-0.42, -0.4]]
std:[[0.02, 0.02, 0.01]][[0.03, 0.03, 0.03]][[0.0, 0.0, 0.0]][[0.01, 0.01]]
MSE:[[0.44, 0.43, 0.43]][[0.59, 0.59, 0.58]][[7.42, 7.42, 7.42]][[0.42, 0.4]]
MSE(-DR):[[0.0, -0.01, -0.01]][[0.15, 0.15, 0.14]][[6.98, 6.98, 6.98]][[-0.02, -0.04]]
better than DR_NO_MARL
MC-based ATE = -0.11
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.06, -0.05, -0.03]][[-0.12, -0.11, -0.12]][[0.11, 0.11, 0.11]][-0.02]
std:[[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.0]
MSE:[[0.06, 0.05, 0.03]][[0.12, 0.11, 0.12]][[0.11, 0.11, 0.11]][0.02]
MSE(-DR):[[0.0, -0.01, -0.03]][[0.06, 0.05, 0.06]][[0.05, 0.05, 0.05]][-0.04]
better than DR_NO_MARL
==============
time spent until now: 14.3 mins


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

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 12.9 21.7 7.1 21.2 6.4

8.5 13.7 11.2 4.3 7.1 15.4

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 1 1 0 1 0

0 1 0 0 0 1

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

1 1 0 1 1 0

0 1 1 0 1 1

1 1 0 1 0 1

1 1 0 1 1 0

1 1 1 0 1 0
```

```
0 1 1 0 0 1

number of reward locations:  23
O_threshold = 15
target policy:

0 0 0 0 0 0

0 0 0 0 0 0

0 0 0 0 0 1

0 0 0 0 0 0

1 0 1 0 1 0

0 0 0 0 0 1

number of reward locations:  5
1
```