```
Last login: Sun Mar 29 22:08:38 on ttys000
Run-Mac:~ mac$ cd ~/.ssh
Run-Mac:.ssh mac$ ssh -i "Runzhe.pem" ubuntu@ec2-3-228-4-227.compute-1.amazonaws.com
Warning: Permanently added the ED25519 host key for IP address '3.228.4.227' 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 02:51:31 UTC 2020

  System load:  0.72               Processes:           224
  Usage of /:   55.4% of 15.45GB   Users logged in:     0
  Memory usage: 1%                 IP address for ens5: 172.31.6.17
  Swap usage:   0%

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

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

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

     https://multipass.run/

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

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-172-31-6-17:~$ export openblas_num_threads=1; export OMP_NUM_THREADS=1ubuntu@ip-172-31-6-17:~$ python EC2.py
22:53, 03/29; num of cores:16

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


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

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

22.3 12.9 16.3 27.0 23.3

7.5 16.1 10.4 10.6 13.0

11.7 19.7 14.9 11.6 13.2

12.6 20.0 10.2 12.5 7.8

4.0 14.3 15.6 8.2 27.3

target policy:

1 1 1 1 1

0 1 0 0 1

0 1 1 0 1

1 1 0 1 0

0 1 1 0 1

number of reward locations:  16
O_threshold = 10
target policy:

1 1 1 1 1

0 1 1 1 1
```

```
1 1 1 1 1

1 1 1 1 0

0 1 1 0 1

number of reward locations:  21
O_threshold = 14
target policy:

1 0 1 1 1

0 1 0 0 0

0 1 1 0 0

0 1 0 0 0

0 1 1 0 1

number of reward locations:  11
^[f1 2 3 1 2 3
--------------------------------------
Value of Behaviour policy:9.684
O_threshold = 12
MC for this TARGET:[10.492, 0.009]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.09, 0.08, 0.05]][[0.14, 0.14, 0.13]][[-10.49, -10.49, -10.49]][[0.05, -0.81]]
std:[[0.03, 0.02, 0.02]][[0.01, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.02, 0.01]]
MSE:[[0.09, 0.08, 0.05]][[0.14, 0.14, 0.13]][[10.49, 10.49, 10.49]][[0.05, 0.81]]
MSE(-DR):[[0.0, -0.01, -0.04]][[0.05, 0.05, 0.04]][[10.4, 10.4, 10.4]][[-0.04, 0.72]]
better than DR_NO_MARL
==============
O_threshold = 10
MC for this TARGET:[10.14, 0.009]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.38, 0.37, 0.35]][[0.41, 0.41, 0.41]][[-10.14, -10.14, -10.14]][[0.35, -0.46]]
std:[[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][[0.0, 0.01]]
MSE:[[0.38, 0.37, 0.35]][[0.41, 0.41, 0.41]][[10.14, 10.14, 10.14]][[0.35, 0.46]]
MSE(-DR):[[0.0, -0.01, -0.03]][[0.03, 0.03, 0.03]][[9.76, 9.76, 9.76]][[-0.03, 0.08]]
better than DR_NO_MARL
MC-based ATE = -0.35
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.29, 0.29, 0.3]][[0.27, 0.27, 0.28]][[0.35, 0.35, 0.35]][0.3]
std:[[0.03, 0.02, 0.02]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.02]
MSE:[[0.29, 0.29, 0.3]][[0.27, 0.27, 0.28]][[0.35, 0.35, 0.35]][0.3]
MSE(-DR):[[0.0, 0.0, 0.01]][[-0.02, -0.02, -0.01]][[0.06, 0.06, 0.06]][0.01]
==============
O_threshold = 14
MC for this TARGET:[10.535, 0.009]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.16, -0.17, -0.17]][[-0.25, -0.25, -0.26]][[-10.54, -10.54, -10.54]][[-0.18, -0.85]]
std:[[0.03, 0.04, 0.01]][[0.0, 0.0, 0.01]][[0.0, 0.0, 0.0]][[0.01, 0.01]]
MSE:[[0.16, 0.17, 0.17]][[0.25, 0.25, 0.26]][[10.54, 10.54, 10.54]][[0.18, 0.85]]
MSE(-DR):[[0.0, 0.01, 0.01]][[0.09, 0.09, 0.1]][[10.38, 10.38, 10.38]][[0.02, 0.69]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
MC-based ATE = 0.04
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.25, -0.25, -0.22]][[-0.39, -0.39, -0.39]][[-0.04, -0.04, -0.04]][-0.23]
std:[[0.06, 0.06, 0.03]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.03]
MSE:[[0.26, 0.26, 0.22]][[0.39, 0.39, 0.39]][[0.04, 0.04, 0.04]][0.23]
MSE(-DR):[[0.0, 0.0, -0.04]][[0.13, 0.13, 0.13]][[-0.22, -0.22, -0.22]][-0.03]
better than DR_NO_MARL
==============
time spent until now: 2.6 mins


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

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

21.1 8.6 8.9 7.2 15.6

4.4 22.2 8.1 12.5 10.0
```

```
19.8 4.8 9.7 9.5 17.3

7.1 10.3 7.8 11.2 13.9

7.1 17.4 15.8 13.5 15.8

target policy:

1 0 0 0 1

0 1 0 1 0

1 0 0 0 1

0 0 0 0 1

0 1 1 1 1

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

1 0 0 0 1

0 1 0 1 0

1 0 0 0 1

0 1 0 1 1

0 1 1 1 1

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

1 0 0 0 1

0 1 0 0 0

1 0 0 0 1

0 0 0 0 0

0 1 1 0 1

number of reward locations:  8
1 MC-based ATE = 0.04
2 3 1 2 3
--------------------------------------
Value of Behaviour policy:7.466
O_threshold = 12
MC for this TARGET:[8.121, 0.009]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.08, -0.09, -0.07]][[-0.18, -0.18, -0.18]][[-8.12, -8.12, -8.12]][[-0.08, -0.65]]
std:[[0.04, 0.04, 0.03]][[0.02, 0.02, 0.02]][[0.0, 0.0, 0.0]][[0.03, 0.0]]
MSE:[[0.09, 0.1, 0.08]][[0.18, 0.18, 0.18]][[8.12, 8.12, 8.12]][[0.09, 0.65]]
MSE(-DR):[[0.0, 0.01, -0.01]][[0.09, 0.09, 0.09]][[8.03, 8.03, 8.03]][[0.0, 0.56]]
better than DR_NO_MARL
==============
O_threshold = 10
MC for this TARGET:[8.077, 0.009]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.06, 0.05, 0.1]][[-0.01, -0.01, -0.02]][[-8.08, -8.08, -8.08]][[0.09, -0.61]]
std:[[0.04, 0.04, 0.03]][[0.02, 0.02, 0.01]][[0.0, 0.0, 0.0]][[0.03, 0.0]]
MSE:[[0.07, 0.06, 0.1]][[0.02, 0.02, 0.02]][[8.08, 8.08, 8.08]][[0.09, 0.61]]
MSE(-DR):[[0.0, -0.01, 0.03]][[-0.05, -0.05, -0.05]][[8.01, 8.01, 8.01]][[0.02, 0.54]]
MC-based ATE = -0.04
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.14, 0.14, 0.17]][[0.16, 0.16, 0.17]][[0.04, 0.04, 0.04]][0.17]
std:[[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.01]
MSE:[[0.14, 0.14, 0.17]][[0.16, 0.16, 0.17]][[0.04, 0.04, 0.04]][0.17]
MSE(-DR):[[0.0, 0.0, 0.03]][[0.02, 0.02, 0.03]][[-0.1, -0.1, -0.1]][0.03]
***** BETTER THAN [IS, DR_NO_MARL] *****
==============
O_threshold = 14
MC for this TARGET:[8.043, 0.009]
```

```
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.23, -0.24, -0.23]][[-0.42, -0.42, -0.43]][[-8.04, -8.04, -8.04]][[-0.24, -0.58]]
std:[[0.03, 0.03, 0.02]][[0.02, 0.02, 0.02]][[0.0, 0.0, 0.0]][[0.02, 0.0]]
MSE:[[0.23, 0.24, 0.23]][[0.42, 0.42, 0.43]][[8.04, 8.04, 8.04]][[0.24, 0.58]]
MSE(-DR):[[0.0, 0.01, 0.0]][[0.19, 0.19, 0.2]][[7.81, 7.81, 7.81]][[0.01, 0.35]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
MC-based ATE = -0.08
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.16, -0.16, -0.17]][[-0.25, -0.25, -0.25]][[0.08, 0.08, 0.08]][-0.16]
std:[[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.01]
MSE:[[0.16, 0.16, 0.17]][[0.25, 0.25, 0.25]][[0.08, 0.08, 0.08]][0.16]
MSE(-DR):[[0.0, 0.0, 0.01]][[0.09, 0.09, 0.09]][[-0.08, -0.08, -0.08]][0.0]
***** BETTER THAN [IS, DR_NO_MARL] *****
==============
time spent until now: 5.0 mins


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

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

9.3 10.8 4.7 21.2 5.4

7.9 13.5 6.7 7.2 7.7

13.7 27.6 11.2 7.0 13.7

8.7 10.9 17.6 8.2 11.1

7.8 10.4 12.2 7.4 9.6

target policy:

0 0 0 1 0

0 1 0 0 0

1 1 0 0 1

0 0 1 0 0

0 0 1 0 0

number of reward locations:  7
O_threshold = 10
target policy:

0 1 0 1 0

0 1 0 0 0

1 1 1 0 1

0 1 1 0 1

0 1 1 0 0

number of reward locations:  12
O_threshold = 14
target policy:

0 0 0 1 0

0 0 0 0 0

0 1 0 0 0

0 0 1 0 0

0 0 0 0 0

number of reward locations:  3
1 2 3 1 2 3
---------------------------------------
Value of Behaviour policy:6.998
```

```
O_threshold = 12
MC for this TARGET:[7.451, 0.008]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.33, -0.33, -0.31]][[-0.44, -0.45, -0.45]][[-7.45, -7.45, -7.45]][[-0.31, -0.45]]
std:[[0.02, 0.02, 0.01]][[0.02, 0.02, 0.02]][[0.0, 0.0, 0.0]][[0.01, 0.0]]
MSE:[[0.33, 0.33, 0.31]][[0.44, 0.45, 0.45]][[7.45, 7.45, 7.45]][[0.31, 0.45]]
MSE(-DR):[[0.0, 0.0, -0.02]][[0.11, 0.12, 0.12]][[7.12, 7.12, 7.12]][[-0.02, 0.12]]
better than DR_NO_MARL
==============
O_threshold = 10
MC for this TARGET:[7.56, 0.008]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.12, -0.12, -0.1]][[-0.08, -0.08, -0.08]][[-7.56, -7.56, -7.56]][[-0.11, -0.56]]
std:[[0.0, 0.01, 0.02]][[0.01, 0.01, 0.0]][[0.0, 0.0, 0.0]][[0.02, 0.0]]
MSE:[[0.12, 0.12, 0.1]][[0.08, 0.08, 0.08]][[7.56, 7.56, 7.56]][[0.11, 0.56]]
MSE(-DR):[[0.0, 0.0, -0.02]][[-0.04, -0.04, -0.04]][[7.44, 7.44, 7.44]][[-0.01, 0.44]]
MC-based ATE = 0.11
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.21, 0.21, 0.21]][[0.36, 0.36, 0.36]][[-0.11, -0.11, -0.11]][0.21]
std:[[0.01, 0.01, 0.01]][[0.02, 0.02, 0.02]][[0.0, 0.0, 0.0]][0.01]
MSE:[[0.21, 0.21, 0.21]][[0.36, 0.36, 0.36]][[0.11, 0.11, 0.11]][0.21]
MSE(-DR):[[0.0, 0.0, 0.0]][[0.15, 0.15, 0.15]][[-0.1, -0.1, -0.1]][0.0]
***** BETTER THAN [IS, DR_NO_MARL] *****
==============
O_threshold = 14
MC for this TARGET:[7.295, 0.008]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[-0.36, -0.36, -0.38]][[-0.65, -0.65, -0.65]][[-7.3, -7.3, -7.3]][[-0.38, -0.3]]
std:[[0.04, 0.04, 0.04]][[0.02, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.03, 0.0]]
MSE:[[0.36, 0.36, 0.38]][[0.65, 0.65, 0.65]][[7.3, 7.3, 7.3]][[0.38, 0.3]]
MSE(-DR):[[0.0, 0.0, 0.02]][[0.29, 0.29, 0.29]][[6.94, 6.94, 6.94]][[0.02, -0.06]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
MC-based ATE = -0.16
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[-0.03, -0.03, -0.07]][[-0.21, -0.2, -0.2]][[0.16, 0.16, 0.16]][-0.07]
std:[[0.02, 0.02, 0.02]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][0.02]
MSE:[[0.04, 0.04, 0.07]][[0.21, 0.2, 0.2]][[0.16, 0.16, 0.16]][0.07]
MSE(-DR):[[0.0, 0.0, 0.03]][[0.17, 0.16, 0.16]][[0.12, 0.12, 0.12]][0.03]
***** BETTER THAN [IS, DR_NO_MARL] *****
==============
time spent until now: 7.4 mins


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

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

22.5 13.1 11.5 5.2 9.9

9.6 10.7 8.6 10.8 9.1

6.5 15.7 15.7 21.8 11.2

9.4 8.9 5.9 16.3 7.1

6.9 10.2 20.0 12.1 7.3

target policy:

1 1 0 0 0

0 0 0 0 0

0 1 1 1 0

0 0 0 1 0

0 0 1 1 0

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

1 1 1 0 0
```

```
0 1 0 1 0

0 1 1 1 1

0 0 0 1 0

0 1 1 1 0

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

1 0 0 0 0

0 0 0 0 0

0 1 1 1 0

0 0 0 1 0

0 0 1 0 0

number of reward locations:  6
```