```
Last login: Sun Mar 29 13:18:09 on ttys000
Run-Mac:~ mac$ cd ~/.ssh
Run-Mac:.ssh mac$ ssh -i "Runzhe.pem" ubuntu@ec2-35-171-129-20.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: Sun Mar 29 17:18:24 2020 from 107.13.161.147
ubuntu@ip-172-31-4-46:~$ export openblas_num_threads=1; export OMP_NUM_THREADS=1
ubuntu@ip-172-31-4-46:~$ python EC2.py
15:26, 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 = 9
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 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 = 11
target policy:

1 1 1 1 1

0 1 0 0 1

1 1 1 1 1

1 1 0 1 0
```

```
0 1 1 0 1

number of reward locations:  18
O_threshold = 13
target policy:

1 0 1 1 1

0 1 0 0 0

0 1 1 0 1

0 1 0 0 0

0 1 1 0 1

number of reward locations:  12
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
^Bd1 2 3 4 1 2 3 4
--------------------------------------
O_threshold = 9
MC-based mean and std of average reward:[1.2473e+01 7.0000e-03]
Value of Behaviour policy:12.208
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.32, 0.31, 0.32]][[0.36, 0.35, 0.36]][[12.47, 12.47, 12.47]][[0.32, 0.26]]
std:[[0.03, 0.03, 0.01]][[0.02, 0.02, 0.03]][[0.0, 0.0, 0.0]][[0.01, 0.01]]
MSE:[[0.32, 0.31, 0.32]][[0.36, 0.35, 0.36]][[12.47, 12.47, 12.47]][[0.32, 0.26]]
MSE(-DR):[[0.0, -0.01, 0.0]][[0.04, 0.03, 0.04]][[12.15, 12.15, 12.15]][[0.0, -0.06]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
==============
O_threshold = 11
MC-based mean and std of average reward:[1.2627e+01 8.0000e-03]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.17, 0.16, 0.16]][[0.19, 0.19, 0.2]][[12.63, 12.63, 12.63]][[0.16, 0.42]]
std:[[0.01, 0.01, 0.0]][[0.02, 0.02, 0.02]][[0.0, 0.0, 0.0]][[0.0, 0.01]]
MSE:[[0.17, 0.16, 0.16]][[0.19, 0.19, 0.2]][[12.63, 12.63, 12.63]][[0.16, 0.42]]
MSE(-DR):[[0.0, -0.01, -0.01]][[0.02, 0.02, 0.03]][[12.46, 12.46, 12.46]][[-0.01, 0.25]]
better than DR_NO_MARL
MC-based ATE = 0.15
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.15, 0.14, 0.16]][[0.16, 0.16, 0.16]][[0.15, 0.15, 0.15]][0.16]
std:[[0.04, 0.04, 0.02]][[0.01, 0.01, 0.0]][[0.0, 0.0, 0.0]][0.01]
MSE:[[0.16, 0.15, 0.16]][[0.16, 0.16, 0.16]][[0.15, 0.15, 0.15]][0.16]
MSE(-DR):[[0.0, -0.01, 0.0]][[0.0, 0.0, 0.0]][[-0.01, -0.01, -0.01]][0.0]
***** BETTER THAN [IS, DR_NO_MARL] *****
==============
O_threshold = 13
MC-based mean and std of average reward:[1.2856e+01 7.0000e-03]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.11, 0.12, 0.09]][[0.14, 0.15, 0.14]][[12.86, 12.86, 12.86]][[0.1, 0.65]]
std:[[0.02, 0.02, 0.02]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.02, 0.01]]
MSE:[[0.11, 0.12, 0.09]][[0.14, 0.15, 0.14]][[12.86, 12.86, 12.86]][[0.1, 0.65]]
MSE(-DR):[[0.0, 0.01, -0.02]][[0.03, 0.04, 0.03]][[12.75, 12.75, 12.75]][[-0.01, 0.54]]
better than DR_NO_MARL
MC-based ATE = 0.38
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.42, 0.43, 0.42]][[0.5, 0.5, 0.5]][[0.38, 0.38, 0.38]][0.42]
std:[[0.05, 0.05, 0.03]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][0.03]
MSE:[[0.42, 0.43, 0.42]][[0.5, 0.5, 0.5]][[0.38, 0.38, 0.38]][0.42]
MSE(-DR):[[0.0, 0.01, 0.0]][[0.08, 0.08, 0.08]][[-0.04, -0.04, -0.04]][0.0]
***** BETTER THAN [IS, DR_NO_MARL] *****
==============
O_threshold = 15
MC-based mean and std of average reward:[1.2905e+01 7.0000e-03]
```

```
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.24, 0.25, 0.2]][[0.35, 0.35, 0.34]][[12.9, 12.9, 12.9]][[0.22, 0.7]]
std:[[0.03, 0.02, 0.01]][[0.01, 0.01, 0.01]][[0.0, 0.0, 0.0]][[0.01, 0.01]]
MSE:[[0.24, 0.25, 0.2]][[0.35, 0.35, 0.34]][[12.9, 12.9, 12.9]][[0.22, 0.7]]
MSE(-DR):[[0.0, 0.01, -0.04]][[0.11, 0.11, 0.1]][[12.66, 12.66, 12.66]][[-0.02, 0.46]]
better than DR_NO_MARL
MC-based ATE = 0.43
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.56, 0.56, 0.53]][[0.7, 0.71, 0.7]][[0.43, 0.43, 0.43]][0.53]
std:[[0.01, 0.01, 0.0]][[0.01, 0.02, 0.02]][[0.0, 0.0, 0.0]][0.01]
MSE:[[0.56, 0.56, 0.53]][[0.7, 0.71, 0.7]][[0.43, 0.43, 0.43]][0.53]
MSE(-DR):[[0.0, 0.0, -0.03]][[0.14, 0.15, 0.14]][[-0.13, -0.13, -0.13]][-0.03]
better than DR_NO_MARL
==============
time spent until now: 3.1 mins


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

max(u_O) =  27.327727595549877
O_threshold = 9
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 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 = 11
target policy:

1 1 1 1 1

0 1 0 0 1

1 1 1 1 1

1 1 0 1 0

0 1 1 0 1

number of reward locations:  18
O_threshold = 13
target policy:

1 0 1 1 1

0 1 0 0 0

0 1 1 0 1

0 1 0 0 0

0 1 1 0 1

number of reward locations:  12
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 4 1 2 3 4
--------------------------------------
O_threshold = 9
MC-based mean and std of average reward:[12.472  0.017]
Value of Behaviour policy:12.212
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.37, 0.37, 0.35]][[0.37, 0.36, 0.37]][[12.47, 12.47, 12.47]][[0.35, 0.26]]
std:[[0.03, 0.03, 0.01]][[0.03, 0.03, 0.03]][[0.0, 0.0, 0.0]][[0.01, 0.01]]
MSE:[[0.37, 0.37, 0.35]][[0.37, 0.36, 0.37]][[12.47, 12.47, 12.47]][[0.35, 0.26]]
MSE(-DR):[[0.0, 0.0, -0.02]][[0.0, -0.01, 0.0]][[12.1, 12.1, 12.1]][[-0.02, -0.11]]
==============
O_threshold = 11
MC-based mean and std of average reward:[12.627  0.017]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.21, 0.21, 0.19]][[0.2, 0.2, 0.2]][[12.63, 12.63, 12.63]][[0.19, 0.42]]
std:[[0.0, 0.0, 0.0]][[0.03, 0.03, 0.03]][[0.0, 0.0, 0.0]][[0.0, 0.01]]
MSE:[[0.21, 0.21, 0.19]][[0.2, 0.2, 0.2]][[12.63, 12.63, 12.63]][[0.19, 0.42]]
MSE(-DR):[[0.0, 0.0, -0.02]][[-0.01, -0.01, -0.01]][[12.42, 12.42, 12.42]][[-0.02, 0.21]]
MC-based ATE = 0.16
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.17, 0.16, 0.16]][[0.17, 0.16, 0.17]][[0.16, 0.16, 0.16]][0.16]
std:[[0.03, 0.03, 0.01]][[0.0, 0.0, 0.01]][[0.0, 0.0, 0.0]][0.01]
MSE:[[0.17, 0.16, 0.16]][[0.17, 0.16, 0.17]][[0.16, 0.16, 0.16]][0.16]
MSE(-DR):[[0.0, -0.01, -0.01]][[0.0, -0.01, 0.0]][[-0.01, -0.01, -0.01]][-0.01]
==============
O_threshold = 13
MC-based mean and std of average reward:[12.856  0.017]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.08, 0.08, 0.06]][[0.11, 0.12, 0.12]][[12.86, 12.86, 12.86]][[0.07, 0.64]]
std:[[0.05, 0.05, 0.03]][[0.03, 0.03, 0.02]][[0.0, 0.0, 0.0]][[0.03, 0.01]]
MSE:[[0.09, 0.09, 0.07]][[0.11, 0.12, 0.12]][[12.86, 12.86, 12.86]][[0.08, 0.64]]
MSE(-DR):[[0.0, 0.0, -0.02]][[0.02, 0.03, 0.03]][[12.77, 12.77, 12.77]][[-0.01, 0.55]]
better than DR_NO_MARL
MC-based ATE = 0.38
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.45, 0.46, 0.41]][[0.48, 0.48, 0.49]][[0.38, 0.38, 0.38]][0.42]
std:[[0.07, 0.07, 0.02]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.02]
MSE:[[0.46, 0.47, 0.41]][[0.48, 0.48, 0.49]][[0.38, 0.38, 0.38]][0.42]
MSE(-DR):[[0.0, 0.01, -0.05]][[0.02, 0.02, 0.03]][[-0.08, -0.08, -0.08]][-0.04]
better than DR_NO_MARL
==============
O_threshold = 15
MC-based mean and std of average reward:[12.904  0.016]
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR/QV/IS]_NO_MF; [DR2, V_behav]
bias:[[0.15, 0.17, 0.15]][[0.32, 0.33, 0.32]][[12.9, 12.9, 12.9]][[0.17, 0.69]]
std:[[0.03, 0.02, 0.01]][[0.03, 0.03, 0.02]][[0.0, 0.0, 0.0]][[0.0, 0.01]]
MSE:[[0.15, 0.17, 0.15]][[0.32, 0.33, 0.32]][[12.9, 12.9, 12.9]][[0.17, 0.69]]
MSE(-DR):[[0.0, 0.02, 0.0]][[0.17, 0.18, 0.17]][[12.75, 12.75, 12.75]][[0.02, 0.54]]
***** BETTER THAN [QV, IS, DR_NO_MARL] *****
MC-based ATE = 0.43
   [DR/QV/IS]; [DR/QV/IS]_NO_MARL; [DR2]
bias:[[0.52, 0.54, 0.5]][[0.69, 0.69, 0.69]][[0.43, 0.43, 0.43]][0.52]
std:[[0.01, 0.0, 0.02]][[0.0, 0.0, 0.0]][[0.0, 0.0, 0.0]][0.01]
MSE:[[0.52, 0.54, 0.5]][[0.69, 0.69, 0.69]][[0.43, 0.43, 0.43]][0.52]
MSE(-DR):[[0.0, 0.02, -0.02]][[0.17, 0.17, 0.17]][[-0.09, -0.09, -0.09]][0.0]
better than DR_NO_MARL
==============
time spent until now: 6.2 mins


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

max(u_O) =  22.15193176791189
O_threshold = 9
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 1

1 0 1 1 1

0 1 0 1 1

0 1 1 1 1

number of reward locations:  16
O_threshold = 11
target policy:

1 0 0 0 1

0 1 0 1 0

1 0 0 0 1

0 0 0 1 1

0 1 1 1 1

number of reward locations:  12
O_threshold = 13
target policy:

1 0 0 0 1

0 1 0 0 0

1 0 0 0 1

0 0 0 0 1

0 1 1 1 1

number of reward locations:  10
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
```