```
Last login: Wed Apr  1 13:03:48 on ttys001
Run-Mac:~ mac$ cd ~/.ssh
Run-Mac:.ssh mac$ ssh -i "Runzhe.pem" ubuntu@ec2-18-204-44-50.compute-1.amazonaws.com
Welcome to Ubuntu 18.04.3 LTS (GNU/Linux 4.15.0-1060-aws x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

  System information as of Wed Apr  1 17:19:15 UTC 2020

  System load:  1.87               Processes:           372
  Usage of /:   56.9% of 15.45GB   Users logged in:     0
  Memory usage: 0%                 IP address for ens5: 172.31.9.82
  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.


*** System restart required ***
Last login: Wed Apr  1 16:43:24 2020 from 107.13.161.147
ubuntu@ip-172-31-9-82:~$ export openblas_num_threads=1; export OMP_NUM_THREADS=1; python EC2.py
13:19, 04/01; num of cores:36

Basic setting:[T, sd_O, sd_D, sd_R, sd_u_O, w_O, w_A,  simple, M_in_R, u_O_u_D, mean_reversion, poisO] = [672, 10, 10, None, 0.3, 0.5, 1
, False, True, 10, False, True]


----------------------------------------
[pattern_seed, sd_R] = [0, 0.5]

max(u_O) =  196.6
O_threshold = 80
means of Order:

168.9 112.2 133.4 194.9 174.2

74.2 132.3 95.1 96.5 112.5

103.9 153.9 125.0 103.2 113.7

110.0 155.7 93.5 109.3 77.0

46.3 121.0 128.9 79.6 196.6

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 = 90
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 = 100
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 = 110
target policy:

1 1 1 1 1

0 1 0 0 1

0 1 1 0 1

0 1 0 0 0

0 1 1 0 1

number of reward locations:  14
O_threshold = 120
target policy:

1 0 1 1 1

0 1 0 0 0

0 1 1 0 0

0 1 0 0 0

0 1 1 0 1

number of reward locations:  11
1 -th target; 2 -th target; 3 -th target; 4 -th target; 5 -th target; one rep DONE
1 -th target; 2 -th target; 3 -th target; 4 -th target; 5 -th target; one rep DONE

----------------------------------------
Value of Behaviour policy:74.592
O_threshold = 80
MC for this TARGET:[84.301, 0.061]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-0.37, -0.71, 0.68]][[4.12, -84.3, -9.71]]
std:[[0.18, 0.18, 0.18]][[0.04, 0.0, 0.01]]
MSE:[[0.41, 0.73, 0.7]][[4.12, 84.3, 9.71]]
MSE(-DR):[[0.0, 0.32, 0.29]][[3.71, 83.89, 9.3]]
***
==============


O_threshold = 90
MC for this TARGET:[84.301, 0.061]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-0.37, -0.71, 0.72]][[4.13, -84.3, -9.71]]
std:[[0.18, 0.18, 0.16]][[0.01, 0.0, 0.01]]
MSE:[[0.41, 0.73, 0.74]][[4.13, 84.3, 9.71]]
MSE(-DR):[[0.0, 0.32, 0.33]][[3.72, 83.89, 9.3]]
***
==============


O_threshold = 100
MC for this TARGET:[89.03, 0.06]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-1.24, -1.67, -2.83]][[0.74, -89.03, -14.44]]
std:[[0.13, 0.12, 0.08]][[0.01, 0.0, 0.01]]
MSE:[[1.25, 1.67, 2.83]][[0.74, 89.03, 14.44]]
MSE(-DR):[[0.0, 0.42, 1.58]][[-0.51, 87.78, 13.19]]
==============


O_threshold = 110
MC for this TARGET:[89.348, 0.06]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-2.97, -3.41, -4.82]][[-2.31, -89.35, -14.76]]
std:[[0.08, 0.03, 0.08]][[0.06, 0.0, 0.01]]
MSE:[[2.97, 3.41, 4.82]][[2.31, 89.35, 14.76]]
MSE(-DR):[[0.0, 0.44, 1.85]][[-0.66, 86.38, 11.79]]
==============


O_threshold = 120
MC for this TARGET:[85.115, 0.06]
```

```
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-2.14, -2.43, -2.46]][[-2.59, -85.12, -10.52]]
std:[[0.23, 0.22, 0.06]][[0.09, 0.0, 0.01]]
MSE:[[2.15, 2.44, 2.46]][[2.59, 85.12, 10.52]]
MSE(-DR):[[0.0, 0.29, 0.31]][[0.44, 82.97, 8.37]]
***
==============


[[ 0.41  0.73  0.7   4.12 84.3   9.71]
 [ 0.41  0.73  0.74  4.13 84.3   9.71]
 [ 1.25  1.67  2.83  0.74 89.03 14.44]
 [ 2.97  3.41  4.82  2.31 89.35 14.76]
 [ 2.15  2.44  2.46  2.59 85.12 10.52]]


time spent until now: 3.2 mins


---------------------------------------
[pattern_seed, sd_R] = [0, 5]

max(u_O) =  196.6
O_threshold = 80
means of Order:

168.9 112.2 133.4 194.9 174.2

74.2 132.3 95.1 96.5 112.5

103.9 153.9 125.0 103.2 113.7

110.0 155.7 93.5 109.3 77.0

46.3 121.0 128.9 79.6 196.6

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 = 90
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 = 100
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 = 110
target policy:

1 1 1 1 1

0 1 0 0 1

0 1 1 0 1

0 1 0 0 0

0 1 1 0 1
```

```
number of reward locations:  14
O_threshold = 120
target policy:

1 0 1 1 1

0 1 0 0 0

0 1 1 0 0

0 1 0 0 0

0 1 1 0 1

number of reward locations:  11
1 -th target; 2 -th target; 3 -th target; 4 -th target; 5 -th target; one rep DONE
1 -th target; 2 -th target; 3 -th target; 4 -th target; 5 -th target; one rep DONE

---------------------------------------
Value of Behaviour policy:74.598
O_threshold = 80
MC for this TARGET:[84.3, 0.065]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-0.13, -0.48, 0.84]][[4.14, -84.3, -9.7]]
std:[[0.22, 0.24, 0.21]][[0.05, 0.0, 0.01]]
MSE:[[0.26, 0.54, 0.87]][[4.14, 84.3, 9.7]]
MSE(-DR):[[0.0, 0.28, 0.61]][[3.88, 84.04, 9.44]]
***
==============


O_threshold = 90
MC for this TARGET:[84.3, 0.065]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-0.13, -0.48, 0.82]][[4.13, -84.3, -9.7]]
std:[[0.2, 0.24, 0.2]][[0.08, 0.0, 0.01]]
MSE:[[0.24, 0.54, 0.84]][[4.13, 84.3, 9.7]]
MSE(-DR):[[0.0, 0.3, 0.6]][[3.89, 84.06, 9.46]]
***
==============


O_threshold = 100
MC for this TARGET:[89.028, 0.068]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-1.23, -1.64, -2.81]][[0.76, -89.03, -14.43]]
std:[[0.08, 0.07, 0.05]][[0.04, 0.0, 0.01]]
MSE:[[1.23, 1.64, 2.81]][[0.76, 89.03, 14.43]]
MSE(-DR):[[0.0, 0.41, 1.58]][[-0.47, 87.8, 13.2]]
==============


O_threshold = 110
MC for this TARGET:[89.346, 0.065]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-3.05, -3.45, -4.93]][[-2.21, -89.35, -14.75]]
std:[[0.08, 0.04, 0.12]][[0.09, 0.0, 0.01]]
MSE:[[3.05, 3.45, 4.93]][[2.21, 89.35, 14.75]]
MSE(-DR):[[0.0, 0.4, 1.88]][[-0.84, 86.3, 11.7]]
==============


O_threshold = 120
MC for this TARGET:[85.113, 0.065]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-2.18, -2.51, -2.5]][[-2.55, -85.11, -10.52]]
std:[[0.26, 0.23, 0.12]][[0.11, 0.0, 0.01]]
MSE:[[2.2, 2.52, 2.5]][[2.55, 85.11, 10.52]]
MSE(-DR):[[0.0, 0.32, 0.3]][[0.35, 82.91, 8.32]]
***
==============


[[ 0.41  0.73  0.7   4.12 84.3   9.71]
 [ 0.41  0.73  0.74  4.13 84.3   9.71]
 [ 1.25  1.67  2.83  0.74 89.03 14.44]
 [ 2.97  3.41  4.82  2.31 89.35 14.76]
 [ 2.15  2.44  2.46  2.59 85.12 10.52]]


[[ 0.26  0.54  0.87  4.14 84.3   9.7 ]
 [ 0.24  0.54  0.84  4.13 84.3   9.7 ]
 [ 1.23  1.64  2.81  0.76 89.03 14.43]
 [ 3.05  3.45  4.93  2.21 89.35 14.75]
 [ 2.2   2.52  2.5   2.55 85.11 10.52]]


time spent until now: 6.5 mins
```

```
---------------------------------------
[pattern_seed, sd_R] = [0, 10]

max(u_O) =  196.6
O_threshold = 80
means of Order:

168.9 112.2 133.4 194.9 174.2

74.2 132.3 95.1 96.5 112.5

103.9 153.9 125.0 103.2 113.7

110.0 155.7 93.5 109.3 77.0

46.3 121.0 128.9 79.6 196.6

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 = 90
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 = 100
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 = 110
target policy:

1 1 1 1 1

0 1 0 0 1

0 1 1 0 1

0 1 0 0 0

0 1 1 0 1

number of reward locations:  14
O_threshold = 120
target policy:

1 0 1 1 1

0 1 0 0 0

0 1 1 0 0

0 1 0 0 0

0 1 1 0 1

number of reward locations:  11
1 -th target; 2 -th target; 3 -th target; 4 -th target; 5 -th target; one rep DONE
```

```
1 -th target; 2 -th target; 3 -th target; 4 -th target; 5 -th target; one rep DONE

---------------------------------------
Value of Behaviour policy:74.606
O_threshold = 80
MC for this TARGET:[84.298, 0.086]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[0.15, -0.23, 0.97]][[4.17, -84.3, -9.69]]
std:[[0.29, 0.32, 0.31]][[0.08, 0.0, 0.01]]
MSE:[[0.33, 0.39, 1.02]][[4.17, 84.3, 9.69]]
MSE(-DR):[[0.0, 0.06, 0.69]][[3.84, 83.97, 9.36]]
***
==============


O_threshold = 90
MC for this TARGET:[84.298, 0.086]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[0.13, -0.23, 0.97]][[4.19, -84.3, -9.69]]
std:[[0.29, 0.32, 0.26]][[0.06, 0.0, 0.01]]
MSE:[[0.32, 0.39, 1.0]][[4.19, 84.3, 9.69]]
MSE(-DR):[[0.0, 0.07, 0.68]][[3.87, 83.98, 9.37]]
***
==============


O_threshold = 100
MC for this TARGET:[89.026, 0.091]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-1.19, -1.6, -2.76]][[0.8, -89.03, -14.42]]
std:[[0.08, 0.02, 0.13]][[0.06, 0.0, 0.01]]
MSE:[[1.19, 1.6, 2.76]][[0.8, 89.03, 14.42]]
MSE(-DR):[[0.0, 0.41, 1.57]][[-0.39, 87.84, 13.23]]
==============


O_threshold = 110
MC for this TARGET:[89.344, 0.086]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-3.05, -3.49, -4.93]][[-2.19, -89.34, -14.74]]
std:[[0.05, 0.05, 0.13]][[0.11, 0.0, 0.01]]
MSE:[[3.05, 3.49, 4.93]][[2.19, 89.34, 14.74]]
MSE(-DR):[[0.0, 0.44, 1.88]][[-0.86, 86.29, 11.69]]
==============


O_threshold = 120
MC for this TARGET:[85.111, 0.086]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-2.23, -2.6, -2.53]][[-2.54, -85.11, -10.51]]
std:[[0.26, 0.25, 0.15]][[0.08, 0.0, 0.01]]
MSE:[[2.25, 2.61, 2.53]][[2.54, 85.11, 10.51]]
MSE(-DR):[[0.0, 0.36, 0.28]][[0.29, 82.86, 8.26]]
***
==============


[[ 0.41  0.73  0.7   4.12 84.3   9.71]
 [ 0.41  0.73  0.74  4.13 84.3   9.71]
 [ 1.25  1.67  2.83  0.74 89.03 14.44]
 [ 2.97  3.41  4.82  2.31 89.35 14.76]
 [ 2.15  2.44  2.46  2.59 85.12 10.52]]


[[ 0.26  0.54  0.87  4.14 84.3   9.7 ]
 [ 0.24  0.54  0.84  4.13 84.3   9.7 ]
 [ 1.23  1.64  2.81  0.76 89.03 14.43]
 [ 3.05  3.45  4.93  2.21 89.35 14.75]
 [ 2.2   2.52  2.5   2.55 85.11 10.52]]


[[ 0.33  0.39  1.02  4.17 84.3   9.69]
 [ 0.32  0.39  1.    4.19 84.3   9.69]
 [ 1.19  1.6   2.76  0.8  89.03 14.42]
 [ 3.05  3.49  4.93  2.19 89.34 14.74]
 [ 2.25  2.61  2.53  2.54 85.11 10.51]]


time spent until now: 9.7 mins


---------------------------------------
[pattern_seed, sd_R] = [0, 20]

max(u_O) =  196.6
O_threshold = 80
means of Order:

168.9 112.2 133.4 194.9 174.2
```

```
74.2 132.3 95.1 96.5 112.5

103.9 153.9 125.0 103.2 113.7

110.0 155.7 93.5 109.3 77.0

46.3 121.0 128.9 79.6 196.6

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 = 90
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 = 100
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 = 110
target policy:

1 1 1 1 1

0 1 0 0 1

0 1 1 0 1

0 1 0 0 0

0 1 1 0 1

number of reward locations:  14
O_threshold = 120
target policy:

1 0 1 1 1

0 1 0 0 0

0 1 1 0 0

0 1 0 0 0

0 1 1 0 1

number of reward locations:  11
1 -th target; 2 -th target; 3 -th target; 4 -th target; 5 -th target; one rep DONE
1 -th target; 2 -th target; 3 -th target; 4 -th target; 5 -th target; one rep DONE

--------------------------------------
Value of Behaviour policy:74.62
O_threshold = 80
MC for this TARGET:[84.294, 0.146]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[0.64, 0.27, 1.23]][[4.24, -84.29, -9.67]]
std:[[0.46, 0.47, 0.52]][[0.16, 0.0, 0.0]]
MSE:[[0.79, 0.54, 1.34]][[4.24, 84.29, 9.67]]
```

```
MSE(-DR):[[0.0, -0.25, 0.55]][[3.45, 83.5, 8.88]]
***
==============


O_threshold = 90
MC for this TARGET:[84.294, 0.146]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[0.66, 0.27, 1.24]][[4.26, -84.29, -9.67]]
std:[[0.42, 0.47, 0.46]][[0.11, 0.0, 0.0]]
MSE:[[0.78, 0.54, 1.32]][[4.26, 84.29, 9.67]]
MSE(-DR):[[0.0, -0.24, 0.54]][[3.48, 83.51, 8.89]]
***
==============


O_threshold = 100
MC for this TARGET:[89.022, 0.152]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-1.09, -1.54, -2.73]][[0.89, -89.02, -14.4]]
std:[[0.06, 0.09, 0.07]][[0.15, 0.0, 0.0]]
MSE:[[1.09, 1.54, 2.73]][[0.9, 89.02, 14.4]]
MSE(-DR):[[0.0, 0.45, 1.64]][[-0.19, 87.93, 13.31]]
==============


O_threshold = 110
MC for this TARGET:[89.34, 0.147]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-3.12, -3.56, -5.0]][[-2.1, -89.34, -14.72]]
std:[[0.09, 0.08, 0.22]][[0.14, 0.0, 0.0]]
MSE:[[3.12, 3.56, 5.0]][[2.1, 89.34, 14.72]]
MSE(-DR):[[0.0, 0.44, 1.88]][[-1.02, 86.22, 11.6]]
==============


O_threshold = 120
MC for this TARGET:[85.107, 0.146]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-2.37, -2.77, -2.62]][[-2.54, -85.11, -10.49]]
std:[[0.32, 0.28, 0.26]][[0.08, 0.0, 0.0]]
MSE:[[2.39, 2.78, 2.63]][[2.54, 85.11, 10.49]]
MSE(-DR):[[0.0, 0.39, 0.24]][[0.15, 82.72, 8.1]]
***
==============


[[ 0.41  0.73  0.7   4.12 84.3   9.71]
 [ 0.41  0.73  0.74  4.13 84.3   9.71]
 [ 1.25  1.67  2.83  0.74 89.03 14.44]
 [ 2.97  3.41  4.82  2.31 89.35 14.76]
 [ 2.15  2.44  2.46  2.59 85.12 10.52]]


[[ 0.26  0.54  0.87  4.14 84.3   9.7 ]
 [ 0.24  0.54  0.84  4.13 84.3   9.7 ]
 [ 1.23  1.64  2.81  0.76 89.03 14.43]
 [ 3.05  3.45  4.93  2.21 89.35 14.75]
 [ 2.2   2.52  2.5   2.55 85.11 10.52]]


[[ 0.33  0.39  1.02  4.17 84.3   9.69]
 [ 0.32  0.39  1.    4.19 84.3   9.69]
 [ 1.19  1.6   2.76  0.8  89.03 14.42]
 [ 3.05  3.49  4.93  2.19 89.34 14.74]
 [ 2.25  2.61  2.53  2.54 85.11 10.51]]


[[ 0.79  0.54  1.34  4.24 84.29  9.67]
 [ 0.78  0.54  1.32  4.26 84.29  9.67]
 [ 1.09  1.54  2.73  0.9  89.02 14.4 ]
 [ 3.12  3.56  5.    2.1  89.34 14.72]
 [ 2.39  2.78  2.63  2.54 85.11 10.49]]


time spent until now: 13.0 mins


---------------------------------------
[pattern_seed, sd_R] = [1, 0.5]

max(u_O) =  167.9
O_threshold = 80
means of Order:

162.0 82.8 84.9 72.1 129.0

49.9 167.9 79.2 109.5 92.3
```

```
154.3 53.6 90.3 88.7 139.8

71.5 94.5 76.5 100.8 118.5

71.5 140.2 130.4 115.7 130.4

target policy:

1 1 1 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:  18
O_threshold = 90
target policy:

1 0 0 0 1

0 1 0 1 1

1 0 1 0 1

0 1 0 1 1

0 1 1 1 1

number of reward locations:  15
O_threshold = 100
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 = 110
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 = 120
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 -th target; 2 -th target; 3 -th target; 4 -th target; 5 -th target; one rep DONE
1 -th target; 2 -th target; 3 -th target; 4 -th target; 5 -th target; one rep DONE

--------------------------------------
Value of Behaviour policy:62.634
O_threshold = 80
MC for this TARGET:[72.685, 0.052]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-1.1, -1.42, -2.07]][[1.79, -72.68, -10.05]]
std:[[0.11, 0.08, 0.17]][[0.02, 0.0, 0.1]]
MSE:[[1.11, 1.42, 2.08]][[1.79, 72.68, 10.05]]
MSE(-DR):[[0.0, 0.31, 0.97]][[0.68, 71.57, 8.94]]
***
==============
```

```
O_threshold = 90
MC for this TARGET:[73.313, 0.046]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-2.74, -3.02, -3.92]][[-0.48, -73.31, -10.68]]
std:[[0.11, 0.11, 0.23]][[0.02, 0.0, 0.1]]
MSE:[[2.74, 3.02, 3.93]][[0.48, 73.31, 10.68]]
MSE(-DR):[[0.0, 0.28, 1.19]][[-2.26, 70.57, 7.94]]
==============


O_threshold = 100
MC for this TARGET:[76.224, 0.049]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-5.76, -6.22, -5.32]][[-4.62, -76.22, -13.59]]
std:[[0.05, 0.07, 0.07]][[0.07, 0.0, 0.1]]
MSE:[[5.76, 6.22, 5.32]][[4.62, 76.22, 13.59]]
MSE(-DR):[[0.0, 0.46, -0.44]][[-1.14, 70.46, 7.83]]
==============


O_threshold = 110
MC for this TARGET:[80.316, 0.052]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-8.74, -9.27, -7.56]][[-9.44, -80.32, -17.68]]
std:[[0.13, 0.13, 0.28]][[0.03, 0.0, 0.1]]
MSE:[[8.74, 9.27, 7.57]][[9.44, 80.32, 17.68]]
MSE(-DR):[[0.0, 0.53, -1.17]][[0.7, 71.58, 8.94]]
**
==============


O_threshold = 120
MC for this TARGET:[81.044, 0.049]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-12.6, -13.08, -10.36]][[-13.58, -81.04, -18.41]]
std:[[0.23, 0.24, 0.37]][[0.01, 0.0, 0.1]]
MSE:[[12.6, 13.08, 10.37]][[13.58, 81.04, 18.41]]
MSE(-DR):[[0.0, 0.48, -2.23]][[0.98, 68.44, 5.81]]
**
==============


[[ 0.41  0.73  0.7   4.12 84.3   9.71]
 [ 0.41  0.73  0.74  4.13 84.3   9.71]
 [ 1.25  1.67  2.83  0.74 89.03 14.44]
 [ 2.97  3.41  4.82  2.31 89.35 14.76]
 [ 2.15  2.44  2.46  2.59 85.12 10.52]]


[[ 0.26  0.54  0.87  4.14 84.3   9.7 ]
 [ 0.24  0.54  0.84  4.13 84.3   9.7 ]
 [ 1.23  1.64  2.81  0.76 89.03 14.43]
 [ 3.05  3.45  4.93  2.21 89.35 14.75]
 [ 2.2   2.52  2.5   2.55 85.11 10.52]]


[[ 0.33  0.39  1.02  4.17 84.3   9.69]
 [ 0.32  0.39  1.    4.19 84.3   9.69]
 [ 1.19  1.6   2.76  0.8  89.03 14.42]
 [ 3.05  3.49  4.93  2.19 89.34 14.74]
 [ 2.25  2.61  2.53  2.54 85.11 10.51]]


[[ 0.79  0.54  1.34  4.24 84.29  9.67]
 [ 0.78  0.54  1.32  4.26 84.29  9.67]
 [ 1.09  1.54  2.73  0.9  89.02 14.4 ]
 [ 3.12  3.56  5.    2.1  89.34 14.72]
 [ 2.39  2.78  2.63  2.54 85.11 10.49]]


[[ 1.11  1.42  2.08  1.79 72.68 10.05]
 [ 2.74  3.02  3.93  0.48 73.31 10.68]
 [ 5.76  6.22  5.32  4.62 76.22 13.59]
 [ 8.74  9.27  7.57  9.44 80.32 17.68]
 [12.6  13.08 10.37 13.58 81.04 18.41]]


time spent until now: 16.2 mins


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

max(u_O) =  167.9
O_threshold = 80
means of Order:
```

```
162.0 82.8 84.9 72.1 129.0

49.9 167.9 79.2 109.5 92.3

154.3 53.6 90.3 88.7 139.8

71.5 94.5 76.5 100.8 118.5

71.5 140.2 130.4 115.7 130.4

target policy:

1 1 1 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:  18
O_threshold = 90
target policy:

1 0 0 0 1

0 1 0 1 1

1 0 1 0 1

0 1 0 1 1

0 1 1 1 1

number of reward locations:  15
O_threshold = 100
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 = 110
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 = 120
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 -th target; 2 -th target; 3 -th target; 4 -th target; 5 -th target; one rep DONE
1 -th target; 2 -th target; 3 -th target; 4 -th target; 5 -th target; one rep DONE

--------------------------------------
Value of Behaviour policy:62.64
O_threshold = 80
MC for this TARGET:[72.683, 0.06]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-1.04, -1.31, -2.04]][[1.8, -72.68, -10.04]]
std:[[0.06, 0.02, 0.04]][[0.02, 0.0, 0.1]]
```

```
MSE:[[1.04, 1.31, 2.04]][[1.8, 72.68, 10.04]]
MSE(-DR):[[0.0, 0.27, 1.0]][[0.76, 71.64, 9.0]]
***
==============


O_threshold = 90
MC for this TARGET:[73.311, 0.056]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-2.69, -3.0, -3.86]][[-0.42, -73.31, -10.67]]
std:[[0.16, 0.12, 0.22]][[0.07, 0.0, 0.1]]
MSE:[[2.69, 3.0, 3.87]][[0.43, 73.31, 10.67]]
MSE(-DR):[[0.0, 0.31, 1.18]][[-2.26, 70.62, 7.98]]
==============


O_threshold = 100
MC for this TARGET:[76.222, 0.062]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-5.64, -6.09, -5.23]][[-4.54, -76.22, -13.58]]
std:[[0.07, 0.09, 0.0]][[0.05, 0.0, 0.1]]
MSE:[[5.64, 6.09, 5.23]][[4.54, 76.22, 13.58]]
MSE(-DR):[[0.0, 0.45, -0.41]][[-1.1, 70.58, 7.94]]
==============


O_threshold = 110
MC for this TARGET:[80.314, 0.066]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-8.68, -9.19, -7.55]][[-9.39, -80.31, -17.67]]
std:[[0.19, 0.17, 0.34]][[0.08, 0.0, 0.1]]
MSE:[[8.68, 9.19, 7.56]][[9.39, 80.31, 17.67]]
MSE(-DR):[[0.0, 0.51, -1.12]][[0.71, 71.63, 8.99]]
**
==============


O_threshold = 120
MC for this TARGET:[81.043, 0.066]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-12.45, -12.93, -10.22]][[-13.49, -81.04, -18.4]]
std:[[0.37, 0.36, 0.51]][[0.08, 0.0, 0.1]]
MSE:[[12.46, 12.94, 10.23]][[13.49, 81.04, 18.4]]
MSE(-DR):[[0.0, 0.48, -2.23]][[1.03, 68.58, 5.94]]
**
==============


[[ 0.41  0.73  0.7   4.12 84.3   9.71]
 [ 0.41  0.73  0.74  4.13 84.3   9.71]
 [ 1.25  1.67  2.83  0.74 89.03 14.44]
 [ 2.97  3.41  4.82  2.31 89.35 14.76]
 [ 2.15  2.44  2.46  2.59 85.12 10.52]]


[[ 0.26  0.54  0.87  4.14 84.3   9.7 ]
 [ 0.24  0.54  0.84  4.13 84.3   9.7 ]
 [ 1.23  1.64  2.81  0.76 89.03 14.43]
 [ 3.05  3.45  4.93  2.21 89.35 14.75]
 [ 2.2   2.52  2.5   2.55 85.11 10.52]]


[[ 0.33  0.39  1.02  4.17 84.3   9.69]
 [ 0.32  0.39  1.    4.19 84.3   9.69]
 [ 1.19  1.6   2.76  0.8  89.03 14.42]
 [ 3.05  3.49  4.93  2.19 89.34 14.74]
 [ 2.25  2.61  2.53  2.54 85.11 10.51]]


[[ 0.79  0.54  1.34  4.24 84.29  9.67]
 [ 0.78  0.54  1.32  4.26 84.29  9.67]
 [ 1.09  1.54  2.73  0.9  89.02 14.4 ]
 [ 3.12  3.56  5.    2.1  89.34 14.72]
 [ 2.39  2.78  2.63  2.54 85.11 10.49]]


[[ 1.11  1.42  2.08  1.79 72.68 10.05]
 [ 2.74  3.02  3.93  0.48 73.31 10.68]
 [ 5.76  6.22  5.32  4.62 76.22 13.59]
 [ 8.74  9.27  7.57  9.44 80.32 17.68]
 [12.6  13.08 10.37 13.58 81.04 18.41]]


[[ 1.04  1.31  2.04  1.8  72.68 10.04]
 [ 2.69  3.    3.87  0.43 73.31 10.67]
 [ 5.64  6.09  5.23  4.54 76.22 13.58]
 [ 8.68  9.19  7.56  9.39 80.31 17.67]
 [12.46 12.94 10.23 13.49 81.04 18.4 ]]
```

time spent until now: 19.4 mins

[pattern_seed, sd_R] = [1, 10]

max(u_O) = 167.9
O_threshold = 80
means of Order:

162.0 82.8 84.9 72.1 129.0

49.9 167.9 79.2 109.5 92.3

154.3 53.6 90.3 88.7 139.8

71.5 94.5 76.5 100.8 118.5

71.5 140.2 130.4 115.7 130.4

target policy:

1 1 1 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: 18
O_threshold = 90
target policy:

1 0 0 0 1

0 1 0 1 1

1 0 1 0 1

0 1 0 1 1

0 1 1 1 1

number of reward locations: 15
O_threshold = 100
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 = 110
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 = 120
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 -th target; 2 -th target; 3 -th target; 4 -th target; 5 -th target; one rep DONE
1 -th target; 2 -th target; 3 -th target; 4 -th target; 5 -th target; one rep DONE

----------------------------------------
Value of Behaviour policy:62.648
O_threshold = 80
MC for this TARGET:[72.681, 0.084]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-0.9, -1.19, -2.04]][[1.8, -72.68, -10.03]]
std:[[0.02, 0.05, 0.1]][[0.08, 0.0, 0.1]]
MSE:[[0.9, 1.19, 2.04]][[1.8, 72.68, 10.03]]
MSE(-DR):[[0.0, 0.29, 1.14]][[0.9, 71.78, 9.13]]
***
==============


O_threshold = 90
MC for this TARGET:[73.309, 0.083]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-2.72, -2.97, -3.92]][[-0.35, -73.31, -10.66]]
std:[[0.15, 0.12, 0.2]][[0.13, 0.0, 0.1]]
MSE:[[2.72, 2.97, 3.93]][[0.37, 73.31, 10.66]]
MSE(-DR):[[0.0, 0.25, 1.21]][[-2.35, 70.59, 7.94]]
==============


O_threshold = 100
MC for this TARGET:[76.22, 0.089]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-5.51, -5.94, -5.17]][[-4.48, -76.22, -13.57]]
std:[[0.1, 0.11, 0.05]][[0.11, 0.0, 0.1]]
MSE:[[5.51, 5.94, 5.17]][[4.48, 76.22, 13.57]]
MSE(-DR):[[0.0, 0.43, -0.34]][[-1.03, 70.71, 8.06]]
==============


O_threshold = 110
MC for this TARGET:[80.312, 0.094]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-8.59, -9.1, -7.42]][[-9.37, -80.31, -17.66]]
std:[[0.25, 0.22, 0.4]][[0.09, 0.0, 0.1]]
MSE:[[8.59, 9.1, 7.43]][[9.37, 80.31, 17.66]]
MSE(-DR):[[0.0, 0.51, -1.16]][[0.78, 71.72, 9.07]]
**
==============


O_threshold = 120
MC for this TARGET:[81.041, 0.095]
   [DR/QV/IS]; [DR_NO_MARL, DR_NO_MF, V_behav]
bias:[[-12.28, -12.77, -10.11]][[-13.43, -81.04, -18.39]]
std:[[0.49, 0.49, 0.64]][[0.13, 0.0, 0.1]]
MSE:[[12.29, 12.78, 10.13]][[13.43, 81.04, 18.39]]
MSE(-DR):[[0.0, 0.49, -2.16]][[1.14, 68.75, 6.1]]
**
==============


[[ 0.41  0.73  0.7   4.12 84.3   9.71]
 [ 0.41  0.73  0.74  4.13 84.3   9.71]
 [ 1.25  1.67  2.83  0.74 89.03 14.44]
 [ 2.97  3.41  4.82  2.31 89.35 14.76]
 [ 2.15  2.44  2.46  2.59 85.12 10.52]]


[[ 0.26  0.54  0.87  4.14 84.3   9.7 ]
 [ 0.24  0.54  0.84  4.13 84.3   9.7 ]
 [ 1.23  1.64  2.81  0.76 89.03 14.43]
 [ 3.05  3.45  4.93  2.21 89.35 14.75]
 [ 2.2   2.52  2.5   2.55 85.11 10.52]]


[[ 0.33  0.39  1.02  4.17 84.3   9.69]
 [ 0.32  0.39  1.    4.19 84.3   9.69]
 [ 1.19  1.6   2.76  0.8  89.03 14.42]
 [ 3.05  3.49  4.93  2.19 89.34 14.74]
 [ 2.25  2.61  2.53  2.54 85.11 10.51]]


[[ 0.79  0.54  1.34  4.24 84.29  9.67]
 [ 0.78  0.54  1.32  4.26 84.29  9.67]
 [ 1.09  1.54  2.73  0.9  89.02 14.4 ]
 [ 3.12  3.56  5.    2.1  89.34 14.72]
 [ 2.39  2.78  2.63  2.54 85.11 10.49]]


[[ 1.11  1.42  2.08  1.79 72.68 10.05]
 [ 2.74  3.02  3.93  0.48 73.31 10.68]
```

```
 [ 5.76  6.22  5.32  4.62 76.22 13.59]
 [ 8.74  9.27  7.57  9.44 80.32 17.68]
 [12.6  13.08 10.37 13.58 81.04 18.41]]


[[ 1.04  1.31  2.04  1.8  72.68 10.04]
 [ 2.69  3.    3.87  0.43 73.31 10.67]
 [ 5.64  6.09  5.23  4.54 76.22 13.58]
 [ 8.68  9.19  7.56  9.39 80.31 17.67]
 [12.46 12.94 10.23 13.49 81.04 18.4 ]]


[[ 0.9   1.19  2.04  1.8  72.68 10.03]
 [ 2.72  2.97  3.93  0.37 73.31 10.66]
 [ 5.51  5.94  5.17  4.48 76.22 13.57]
 [ 8.59  9.1   7.43  9.37 80.31 17.66]
 [12.29 12.78 10.13 13.43 81.04 18.39]]


time spent until now: 22.7 mins


---------------------------------------
[pattern_seed, sd_R] = [1, 20]

max(u_O) =  167.9
O_threshold = 80
means of Order:

162.0 82.8 84.9 72.1 129.0

49.9 167.9 79.2 109.5 92.3

154.3 53.6 90.3 88.7 139.8

71.5 94.5 76.5 100.8 118.5

71.5 140.2 130.4 115.7 130.4

target policy:

1 1 1 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:  18
O_threshold = 90
target policy:

1 0 0 0 1

0 1 0 1 1

1 0 1 0 1

0 1 0 1 1

0 1 1 1 1

number of reward locations:  15
O_threshold = 100
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 = 110
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
Q_threshold = 120
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 -th target; 2 -th target;
```