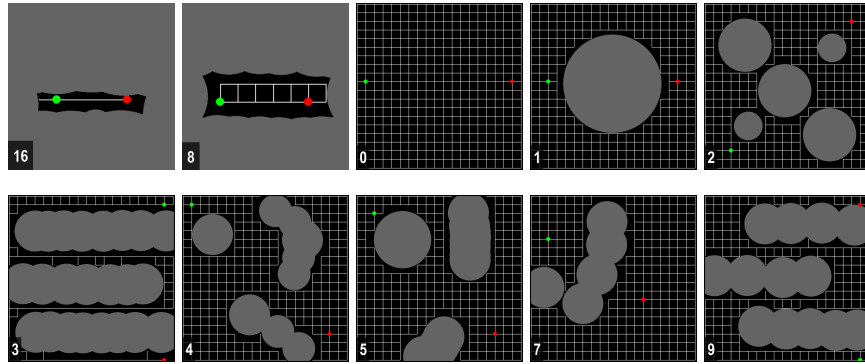
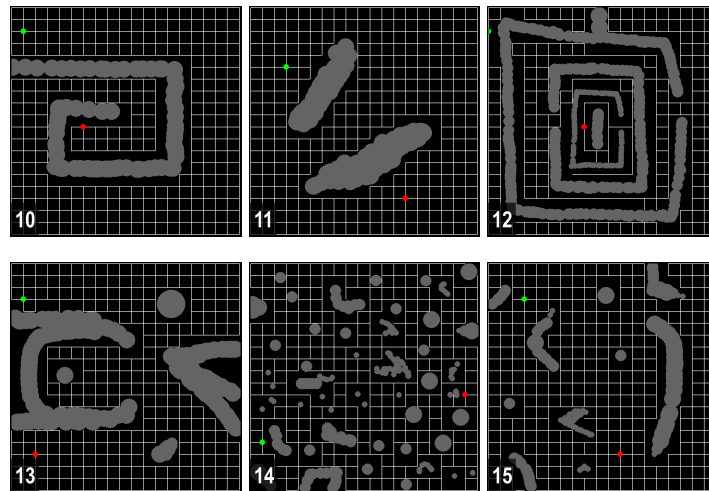


### A Problems



**Fig. 4.** Ten problems (enumerated from top-left to bottom-right as 16, 8, 0, 1, 2, 3, 4, 5, 7, 9). The green dot is the initial state, while the red dot is the goal state. The white lines connect the states and can be viewed as the actions allowed by the robot. The gray circles represent obstacles that cannot be passed through



**Fig. 5.** Additional six problems (enumerated from top-left to bottom-right as 10 - 15).

## B Tables for Deterministic Experiments

The performance for Q-learning and the dynamic programming methods on deterministic problems, run on a desktop computer with a Ryzen 7 3700X CPU (3.6GHz), is shown in the following tables. We show the run times of the algorithms and the number of actions that were used during that run time. We also show how often the algorithm converges and how long it took for its exploration policy to discover the goal for the first time, together with the time it took for the initial state to achieve its optimal cost to go value, tracking the mean and standard deviation for each applicable metric. The last three tables of this subsection show the same results for Problems 1, 10 and 11 run on a desktop computer with an i5-12400F CPU (2.50GHz), but also including a row for the performance of Q-learning when using  $\epsilon$ -decay.

Algorithm (Problem 0)	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)
Q-learning ( $\epsilon = 0$ )	0.11326 $\pm$ 0.0007	32774.0000 $\pm$ 0.0000	0.0 %	0.0075 $\pm$ 0.0000	2131.0000 $\pm$ 0.0000	0.0622 $\pm$ 0.0003	18000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 0.25$ )	0.15533 $\pm$ 0.0019	45186.3700 $\pm$ 270.0393	0.0 %	0.0138 $\pm$ 0.0019	4041.3800 $\pm$ 537.1843	0.0691 $\pm$ 0.0014	20140.0000 $\pm$ 374.6999
Q-learning ( $\epsilon = 0.5$ )	0.24147 $\pm$ 0.0025	72453.8400 $\pm$ 618.0135	0.0 %	0.0181 $\pm$ 0.0029	5436.8200 $\pm$ 871.1261	0.0831 $\pm$ 0.0023	24990.0000 $\pm$ 655.6676
Q-learning ( $\epsilon = 0.75$ )	0.58168 $\pm$ 0.0079	180569.8300 $\pm$ 2367.9754	0.0 %	0.0199 $\pm$ 0.0048	6167.2200 $\pm$ 1481.6033	0.1117 $\pm$ 0.0038	34790.0000 $\pm$ 1210.7436
Q-learning ( $\epsilon = 0.9$ )	0.95443 $\pm$ 0.0164	303767.8100 $\pm$ 52175.4497	100.0 %	0.0158 $\pm$ 0.0081	5039.6800 $\pm$ 2570.6394	0.1504 $\pm$ 0.0076	48030.0000 $\pm$ 2451.3466
Q-learning ( $\epsilon = 1$ )	0.40398 $\pm$ 0.0332	135913.8400 $\pm$ 11113.1169	100.0 %	0.0050 $\pm$ 0.0039	1636.8800 $\pm$ 1303.7126	0.2082 $\pm$ 0.0192	69980.0000 $\pm$ 6452.8753
Q-learning ( $\pi$ ) ( $\epsilon = 1$ )	0.41524 $\pm$ 0.0313	136706.6400 $\pm$ 10296.1120	100.0 %	0.0042 $\pm$ 0.0034	1359.0800 $\pm$ 1101.2584	0.2114 $\pm$ 0.0187	69530.0000 $\pm$ 6119.5670
Model-free Dijkstra	0.00295 $\pm$ 0.0005	5368.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free							
Asynchronous Value Iteration	0.02134 $\pm$ 0.0005	31654.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Value Iteration	0.03276 $\pm$ 0.0007	48374.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan

Algorithm (Problem 1)	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)
Q-learning ( $\epsilon = 0$ )	0.13929 $\pm$ 0.0009	40894.0000 $\pm$ 0.0000	0.0 %	0.0039 $\pm$ 0.0000	1127.0000 $\pm$ 0.0000	0.0688 $\pm$ 0.0005	20000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 0.25$ )	0.18457 $\pm$ 0.0016	53343.1400 $\pm$ 228.6908	0.0 %	0.0074 $\pm$ 0.0025	2141.1200 $\pm$ 731.3317	0.0817 $\pm$ 0.0018	23570.0000 $\pm$ 495.0758
Q-learning ( $\epsilon = 0.5$ )	0.26569 $\pm$ 0.0025	79250.9600 $\pm$ 530.5786	0.0 %	0.0096 $\pm$ 0.0031	2861.3800 $\pm$ 926.8256	0.0984 $\pm$ 0.0025	29390.0000 $\pm$ 719.6527
Q-learning ( $\epsilon = 0.75$ )	0.54417 $\pm$ 0.0282	172628.4100 $\pm$ 8373.3798	9.0 %	0.0111 $\pm$ 0.0037	3518.6200 $\pm$ 1163.6424	0.1321 $\pm$ 0.0058	42070.0000 $\pm$ 1779.0728
Q-learning ( $\epsilon = 0.9$ )	0.28439 $\pm$ 0.0376	92948.4700 $\pm$ 12287.7324	100.0 %	0.0102 $\pm$ 0.0064	3342.1800 $\pm$ 2101.8655	0.1823 $\pm$ 0.0126	59540.0000 $\pm$ 4016.0179
Q-learning ( $\epsilon = 1$ )	0.33558 $\pm$ 0.0396	114579.3300 $\pm$ 13500.2663	100.0 %	0.0045 $\pm$ 0.0040	1540.3600 $\pm$ 1391.8213	0.2837 $\pm$ 0.0396	96820.0000 $\pm$ 13717.4196
Q-learning ( $\pi$ ) ( $\epsilon = 1$ )	0.34452 $\pm$ 0.0405	114637.3700 $\pm$ 13326.4214	100.0 %	0.0050 $\pm$ 0.0042	1643.4800 $\pm$ 1398.7443	0.2925 $\pm$ 0.0388	97310.0000 $\pm$ 12780.2152
Model-free Dijkstra	0.00224 $\pm$ 0.0005	4015.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free							
Asynchronous Value Iteration	0.01638 $\pm$ 0.0006	23308.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Value Iteration	0.02401 $\pm$ 0.0007	35078.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan

Algorithm (Problem 2)	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)
Q-learning ( $\epsilon = 0$ )	0.14180 $\pm$ 0.0012	43272.0000 $\pm$ 0.0000	0.0 %	0.0030 $\pm$ 0.0000	905.0000 $\pm$ 0.0000	0.0662 $\pm$ 0.0004	20000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 0.25$ )	0.18772 $\pm$ 0.0014	57044.9800 $\pm$ 230.5382	0.0 %	0.0051 $\pm$ 0.0022	1551.0000 $\pm$ 655.2924	0.0793 $\pm$ 0.0014	24020.0000 $\pm$ 399.4997
Q-learning ( $\epsilon = 0.5$ )	0.27709 $\pm$ 0.0047	86112.3600 $\pm$ 627.8763	0.0 %	0.0076 $\pm$ 0.0027	2356.4200 $\pm$ 837.0154	0.0963 $\pm$ 0.0027	29990.0000 $\pm$ 768.0495
Q-learning ( $\epsilon = 0.75$ )	0.60526 $\pm$ 0.0143	194341.4700 $\pm$ 4266.8579	4.0 %	0.0094 $\pm$ 0.0040	3016.5000 $\pm$ 1293.1793	0.1322 $\pm$ 0.0056	42610.0000 $\pm$ 1719.8546
Q-learning ( $\epsilon = 0.9$ )	0.31286 $\pm$ 0.0407	103589.5700 $\pm$ 13507.4629	100.0 %	0.0099 $\pm$ 0.0064	3264.3800 $\pm$ 2105.1116	0.1840 $\pm$ 0.0130	60960.0000 $\pm$ 4300.9766
Q-learning ( $\epsilon = 1$ )	0.44531 $\pm$ 0.0903	153230.1400 $\pm$ 30413.1369	100.0 %	0.0069 $\pm$ 0.0061	2365.7600 $\pm$ 2124.2876	0.3170 $\pm$ 0.0630	109180.0000 $\pm$ 21751.9562
Q-learning ( $\pi$ ) ( $\epsilon = 1$ )	0.45306 $\pm$ 0.0842	150127.5800 $\pm$ 27994.8262	100.0 %	0.0079 $\pm$ 0.0059	2601.4600 $\pm$ 1943.0290	0.3359 $\pm$ 0.0605	111290.0000 $\pm$ 19900.3995
Model-free Dijkstra	0.00208 $\pm$ 0.0000	3665.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free							
Asynchronous Value Iteration	0.02007 $\pm$ 0.0007	27660.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Value Iteration	0.02591 $\pm$ 0.0006	37640.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan

Algorithm (Problem 3)	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)
Q-learning ( $\epsilon = 0$ )	0.31918 $\pm$ 0.0040	104562.0000 $\pm$ 0.0000	0.0 %	0.0049 $\pm$ 0.0001	1602.0000 $\pm$ 0.0000	0.0792 $\pm$ 0.0012	26000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 0.25$ )	0.13560 $\pm$ 0.0222	43657.4800 $\pm$ 7142.5875	100.0 %	0.0054 $\pm$ 0.0033	1733.7200 $\pm$ 1060.4745	0.0958 $\pm$ 0.0030	30940.0000 $\pm$ 822.4354
Q-learning ( $\epsilon = 0.5$ )	0.13557 $\pm$ 0.0088	44921.7400 $\pm$ 2909.6461	100.0 %	0.0074 $\pm$ 0.0046	2452.8400 $\pm$ 1545.4950	0.1175 $\pm$ 0.0039	38980.0000 $\pm$ 1224.5816
Q-learning ( $\epsilon = 0.75$ )	0.17153 $\pm$ 0.0095	58706.1600 $\pm$ 3219.7965	100.0 %	0.0126 $\pm$ 0.0069	4289.3400 $\pm$ 2352.6951	0.1590 $\pm$ 0.0077	54460.0000 $\pm$ 2566.7879
Q-learning ( $\epsilon = 0.9$ )	0.22911 $\pm$ 0.0148	79609.0600 $\pm$ 5036.7241	100.0 %	0.0238 $\pm$ 0.0140	8292.6800 $\pm$ 4890.7428	0.2145 $\pm$ 0.0137	74610.0000 $\pm$ 4780.9936
Q-learning ( $\epsilon = 1$ )	0.99921 $\pm$ 0.3629	365387.0000 $\pm$ 132113.6869	100.0 %	0.1230 $\pm$ 0.1161	45427.7600 $\pm$ 42713.2396	0.9842 $\pm$ 0.3651	359940.0000 $\pm$ 132921.3166
Q-learning ( $\pi$ ) ( $\epsilon = 1$ )	0.97154 $\pm$ 0.2853	334139.2400 $\pm$ 97320.2744	100.0 %	0.1417 $\pm$ 0.1139	49120.8400 $\pm$ 38980.7772	0.9630 $\pm$ 0.2830	331240.0000 $\pm$ 96539.8488
Model-free Dijkstra	0.00069 $\pm$ 0.0000	1061.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free							
Asynchronous Value Iteration	0.01782 $\pm$ 0.0006	23774.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Value Iteration	0.02389 $\pm$ 0.0009	31790.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan

Algorithm (Problem 4)	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)
Q-learning ( $\epsilon = 0$ )	0.15125 $\pm$ 0.0034	46110.0000 $\pm$ 0.0000	0.0 %	0.0044 $\pm$ 0.0003	1325.0000 $\pm$ 0.0000	0.0724 $\pm$ 0.0021	22000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 0.25$ )	0.19840 $\pm$ 0.0022	60374.1400 $\pm$ 315.4138	0.0 %	0.0147 $\pm$ 0.0037	4472.6600 $\pm$ 1134.9649	0.0869 $\pm$ 0.0021	26480.0000 $\pm$ 538.1450
Q-learning ( $\epsilon = 0.5$ )	0.28498 $\pm$ 0.0053	88913.2900 $\pm$ 649.2647	0.0 %	0.0177 $\pm$ 0.0043	5566.3400 $\pm$ 1348.1139	0.1081 $\pm$ 0.0039	33890.0000 $\pm$ 1038.2196
Q-learning ( $\epsilon = 0.75$ )	0.57889 $\pm$ 0.0082	186567.6800 $\pm$ 2407.6892	0.0 %	0.0203 $\pm$ 0.0061	6607.7800 $\pm$ 1982.5550	0.1520 $\pm$ 0.0090	49330.0000 $\pm$ 2956.5351
Q-learning ( $\epsilon = 0.9$ )	1.36798 $\pm$ 0.2875	451724.6400 $\pm$ 94716.2085	58.0 %	0.0168 $\pm$ 0.0098	5610.9200 $\pm$ 3275.1360	0.2096 $\pm$ 0.0166	69800.0000 $\pm$ 5499.0908
Q-learning ( $\epsilon = 1$ )	0.80289 $\pm$ 0.1904	278558.3600 $\pm$ 65909.6114	100.0 %	0.0085 $\pm$ 0.0086	2930.4200 $\pm$ 2986.4306	0.3581 $\pm$ 0.0701	124770.0000 $\pm$ 24473.2866
Q-learning ( $\pi$ ) ( $\epsilon = 1$ )	0.81879 $\pm$ 0.2648	277054.3300 $\pm$ 88885.4437	100.0 %	0.0076 $\pm$ 0.0073	2578.1200 $\pm$ 2140.4350	0.3647 $\pm$ 0.0628	124020.0000 $\pm$ 21313.5378
Model-free Dijkstra	0.00207 $\pm$ 0.0005	3663.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Asynchronous Value Iteration	0.01690 $\pm$ 0.0005	24447.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Value Iteration	0.02575 $\pm$ 0.0005	37837.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan

Algorithm (Problem 5)	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)
Q-learning ( $\epsilon = 0$ )	0.14089 $\pm$ 0.0011	42386.0000 $\pm$ 0.0000	0.0 %	0.0047 $\pm$ 0.0001	1430.0000 $\pm$ 0.0000	0.0625 $\pm$ 0.0004	19000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 0.25$ )	0.18808 $\pm$ 0.0012	56130.6800 $\pm$ 264.3807	0.0 %	0.0107 $\pm$ 0.0029	3236.7400 $\pm$ 879.1809	0.0746 $\pm$ 0.0019	22440.0000 $\pm$ 571.3143
Q-learning ( $\epsilon = 0.5$ )	0.27472 $\pm$ 0.0027	84451.7000 $\pm$ 702.8104	0.0 %	0.0139 $\pm$ 0.0029	4341.1200 $\pm$ 894.5730	0.0940 $\pm$ 0.0044	29150.0000 $\pm$ 1336.9742
Q-learning ( $\epsilon = 0.75$ )	0.58191 $\pm$ 0.0116	184877.3400 $\pm$ 2306.1895	0.0 %	0.0152 $\pm$ 0.0044	4888.6000 $\pm$ 1401.3910	0.1350 $\pm$ 0.0120	43400.0000 $\pm$ 3778.8887
Q-learning ( $\epsilon = 0.9$ )	0.98171 $\pm$ 0.2661	319656.4200 $\pm$ 86716.4391	98.0 %	0.0112 $\pm$ 0.0065	3673.4600 $\pm$ 2140.2486	0.1834 $\pm$ 0.0206	60150.0000 $\pm$ 6768.7775
Q-learning ( $\epsilon = 1$ )	0.49500 $\pm$ 0.1111	170925.8200 $\pm$ 38202.6919	100.0 %	0.0058 $\pm$ 0.0048	1977.1400 $\pm$ 1657.1752	0.2829 $\pm$ 0.0489	97870.0000 $\pm$ 16768.8133
Q-learning ( $\pi$ ) ( $\epsilon = 1$ )	0.53800 $\pm$ 0.1353	179624.1400 $\pm$ 44847.8988	100.0 %	0.0060 $\pm$ 0.0046	2018.1600 $\pm$ 1544.2775	0.2987 $\pm$ 0.0459	100000.0000 $\pm$ 15316.0047
Model-free Dijkstra	0.00204 $\pm$ 0.0002	3615.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Asynchronous Value Iteration	0.01620 $\pm$ 0.0007	23059.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Value Iteration	0.02471 $\pm$ 0.0006	36241.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan

Algorithm (Problem 7)	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)
Q-learning ( $\epsilon = 0$ )	0.12079 $\pm$ 0.0007	37478.0000 $\pm$ 0.0000	0.0 %	0.0029 $\pm$ 0.0000	880.0000 $\pm$ 0.0000	0.0459 $\pm$ 0.0004	14000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 0.25$ )	0.16239 $\pm$ 0.0016	49680.7600 $\pm$ 323.7531	0.0 %	0.0050 $\pm$ 0.0018	1508.2000 $\pm$ 544.4474	0.0538 $\pm$ 0.0015	16270.0000 $\pm$ 443.9595
Q-learning ( $\epsilon = 0.5$ )	0.24275 $\pm$ 0.0027	76035.4800 $\pm$ 661.8601	0.0 %	0.0068 $\pm$ 0.0023	2071.0600 $\pm$ 695.5948	0.0669 $\pm$ 0.0022	20780.0000 $\pm$ 672.0119
Q-learning ( $\epsilon = 0.75$ )	0.52271 $\pm$ 0.0068	169471.7200 $\pm$ 2123.6347	0.0 %	0.0087 $\pm$ 0.0036	2803.9400 $\pm$ 1157.2607	0.0927 $\pm$ 0.0042	29970.0000 $\pm$ 1374.4453
Q-learning ( $\epsilon = 0.9$ )	1.50883 $\pm$ 0.1231	499216.7200 $\pm$ 40690.9169	34.0 %	0.0077 $\pm$ 0.0047	2541.0600 $\pm$ 1558.5223	0.1312 $\pm$ 0.0111	43450.0000 $\pm$ 3691.5444
Q-learning ( $\epsilon = 1$ )	1.82401 $\pm$ 0.3753	633100.6200 $\pm$ 130431.7955	100.0 %	0.0081 $\pm$ 0.0078	2820.5200 $\pm$ 2756.1423	0.3153 $\pm$ 0.0780	109720.0000 $\pm$ 27010.0278
Q-learning ( $\pi$ ) ( $\epsilon = 1$ )	1.96900 $\pm$ 0.6300	662812.2800 $\pm$ 211950.3269	100.0 %	0.0095 $\pm$ 0.0107	3198.5800 $\pm$ 3640.9636	0.3005 $\pm$ 0.0670	101600.0000 $\pm$ 22493.5546
Model-free Dijkstra	0.00217 $\pm$ 0.0005	3856.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Asynchronous Value Iteration	0.01852 $\pm$ 0.0005	26437.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Value Iteration	0.02888 $\pm$ 0.0001	42649.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan

Algorithm (Problem 8)	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)
Q-learning ( $\epsilon = 0$ )	0.01600 $\pm$ 0.0002	5094.0000 $\pm$ 0.0000	0.0 %	0.0001 $\pm$ 0.0000	43.0000 $\pm$ 0.0000	0.0032 $\pm$ 0.0001	1000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 0.25$ )	0.02015 $\pm$ 0.0051	6466.8900 $\pm$ 1658.8178	26.0 %	0.0002 $\pm$ 0.0000	51.9400 $\pm$ 9.1933	0.0032 $\pm$ 0.0000	1000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 0.5$ )	0.01306 $\pm$ 0.0078	4306.3500 $\pm$ 2580.3645	98.0 %	0.0002 $\pm$ 0.0001	59.1200 $\pm$ 17.9840	0.0031 $\pm$ 0.0000	1000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 0.75$ )	0.00628 $\pm$ 0.0033	2127.4100 $\pm$ 1121.6373	100.0 %	0.0002 $\pm$ 0.0001	60.1200 $\pm$ 28.4904	0.0030 $\pm$ 0.0000	1000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 0.9$ )	0.00510 $\pm$ 0.0025	1760.2800 $\pm$ 867.9923	100.0 %	0.0002 $\pm$ 0.0001	54.9200 $\pm$ 36.0288	0.0030 $\pm$ 0.0000	1000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 1$ )	0.00458 $\pm$ 0.0020	1655.8500 $\pm$ 736.9191	100.0 %	0.0001 $\pm$ 0.0001	40.5600 $\pm$ 31.1019	0.0028 $\pm$ 0.0000	1000.0000 $\pm$ 0.0000
Q-learning ( $\pi$ ) ( $\epsilon = 1$ )	0.00591 $\pm$ 0.0024	2022.0400 $\pm$ 856.0125	100.0 %	0.0001 $\pm$ 0.0001	43.6600 $\pm$ 32.3469	0.0030 $\pm$ 0.0001	1000.0000 $\pm$ 0.0000
Model-free Dijkstra	0.00008 $\pm$ 0.0000	119.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Asynchronous Value Iteration	0.00025 $\pm$ 0.0000	319.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Value Iteration	0.00027 $\pm$ 0.0000	357.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan

Algorithm (Problem 9)	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)
Q-learning ( $\epsilon = 0$ )	0.30094 $\pm$ 0.0025	93512.0000 $\pm$ 0.0000	0.0 %	0.0030 $\pm$ 0.0000	924.0000 $\pm$ 0.0000	0.1071 $\pm$ 0.0010	33000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 0.25$ )	0.40332 $\pm$ 0.0026	124056.6000 $\pm$ 369.5486	0.0 %	0.0063 $\pm$ 0.0049	1920.4400 $\pm$ 1473.8903	0.1266 $\pm$ 0.0022	38730.0000 $\pm$ 580.6031
Q-learning ( $\epsilon = 0.5$ )	0.59771 $\pm$ 0.0045	189069.5400 $\pm$ 805.8806	0.0 %	0.0103 $\pm$ 0.0063	3219.8200 $\pm$ 1985.7043	0.1528 $\pm$ 0.0037	48240.0000 $\pm$ 1132.4310
Q-learning ( $\epsilon = 0.75$ )	0.32806 $\pm$ 0.0497	106893.5400 $\pm$ 16033.8762	100.0 %	0.0160 $\pm$ 0.0085	5189.9200 $\pm$ 2773.5873	0.2082 $\pm$ 0.0080	67810.0000 $\pm$ 2561.4644
Q-learning ( $\epsilon = 0.9$ )	0.30673 $\pm$ 0.0195	102077.7200 $\pm$ 6343.1919	100.0 %	0.0330 $\pm$ 0.0177	10934.4600 $\pm$ 5863.3670	0.2969 $\pm$ 0.0170	98810.0000 $\pm$ 5556.4287
Q-learning ( $\epsilon = 1$ )	1.74662 $\pm$ 0.7671	611856.5000 $\pm$ 269104.1334	100.0 %	0.1213 $\pm$ 0.1074	42628.7200 $\pm$ 37559.9343	1.2952 $\pm$ 0.4267	453590.0000 $\pm$ 149158.4456
Q-learning ( $\pi$ ) ( $\epsilon = 1$ )	2.10442 $\pm$ 1.1306	704462.5600 $\pm$ 377424.9032	100.0 %	0.1779 $\pm$ 0.1123	59863.2600 $\pm$ 37681.3580	1.3004 $\pm$ 0.2590	435550.0000 $\pm$ 87139.9306
Model-free Dijkstra	0.00143 $\pm$ 0.0005	2360.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Asynchronous Value Iteration	0.02774 $\pm$ 0.0002	38776.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Value Iteration	0.03483 $\pm$ 0.0002	50692.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan

Algorithm (Problem 10)	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)
Q-learning ( $\epsilon = 0$ )	0.33383 $\pm$ 0.0025	101214.0000 $\pm$ 0.0000	0.0 %	0.0068 $\pm$ 0.0002	2015.0000 $\pm$ 0.0000	0.1720 $\pm$ 0.0016	52000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 0.25$ )	0.44088 $\pm$ 0.0030	133171.4600 $\pm$ 405.0955	0.0 %	0.0101 $\pm$ 0.0058	3050.0200 $\pm$ 1747.4311	0.2043 $\pm$ 0.0041	61830.0000 $\pm$ 1200.4582
Q-learning ( $\epsilon = 0.5$ )	0.64394 $\pm$ 0.0045	200028.6600 $\pm$ 996.3678	0.0 %	0.0168 $\pm$ 0.0071	5222.4200 $\pm$ 2184.3765	0.2517 $\pm$ 0.0066	78540.0000 $\pm$ 1951.5122
Q-learning ( $\epsilon = 0.75$ )	1.08726 $\pm$ 0.2099	348322.6400 $\pm$ 66374.2689	76.0 %	0.0231 $\pm$ 0.0110	7449.8000 $\pm$ 3206.2196	0.3502 $\pm$ 0.0129	112990.0000 $\pm$ 3410.2639
Q-learning ( $\epsilon = 0.9$ )	0.61889 $\pm$ 0.0559	205682.0500 $\pm$ 18488.2891	100.0 %	0.0372 $\pm$ 0.0194	12384.0200 $\pm$ 6398.5418	0.4858 $\pm$ 0.0239	161590.0000 $\pm$ 7676.8776
Q-learning ( $\epsilon = 1$ )	1.88356 $\pm$ 0.6740	657763.2100 $\pm$ 235315.6482	100.0 %	0.0860 $\pm$ 0.0912	30118.4200 $\pm$ 32113.6083	1.2836 $\pm$ 0.3801	448790.0000 $\pm$ 133803.4600
Q-learning ( $\pi$ ) ( $\epsilon = 1$ )	2.02001 $\pm$ 0.4996	685022.7200 $\pm$ 169670.6191	100.0 %	0.0539 $\pm$ 0.0569	18366.3400 $\pm$ 19499.6667	1.3425 $\pm$	

Algorithm (Problem 11)	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)
Q-learning ( $\epsilon = 0$ )	0.11746 $\pm$ 0.0010	35316.0000 $\pm$ 0.0000	0.0 %	0.0052 $\pm$ 0.0001	1569.0000 $\pm$ 0.0000	0.0562 $\pm$ 0.0005	17000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 0.25$ )	0.16348 $\pm$ 0.0017	48929.1500 $\pm$ 308.7902	0.0 %	0.0069 $\pm$ 0.0019	2083.6000 $\pm$ 575.2788	0.0674 $\pm$ 0.0018	20340.0000 $\pm$ 533.2917
Q-learning ( $\epsilon = 0.5$ )	0.24302 $\pm$ 0.0030	75008.7100 $\pm$ 538.2311	0.0 %	0.0083 $\pm$ 0.0028	2588.2400 $\pm$ 895.4427	0.0801 $\pm$ 0.0022	24950.0000 $\pm$ 683.7397
Q-learning ( $\epsilon = 0.75$ )	0.52179 $\pm$ 0.0074	166191.9100 $\pm$ 2022.8075	0.0 %	0.0099 $\pm$ 0.0041	3199.6400 $\pm$ 1346.4165	0.1070 $\pm$ 0.0044	34520.0000 $\pm$ 1403.4244
Q-learning ( $\epsilon = 0.9$ )	0.83570 $\pm$ 0.0194	272845.6200 $\pm$ 61283.9929	99.0 %	0.0093 $\pm$ 0.0061	3071.7600 $\pm$ 2019.3443	0.1416 $\pm$ 0.0071	46750.0000 $\pm$ 2329.6996
Q-learning ( $\epsilon = 1$ )	0.41468 $\pm$ 0.0594	144164.0500 $\pm$ 20481.0098	100.0 %	0.0054 $\pm$ 0.0041	1859.9000 $\pm$ 1424.8237	0.2162 $\pm$ 0.0341	75080.0000 $\pm$ 11847.0925
Q-learning ( $\pi$ ) ( $\epsilon = 1$ )	0.42675 $\pm$ 0.0642	144673.6700 $\pm$ 21823.7557	100.0 %	0.0055 $\pm$ 0.0055	1936.9800 $\pm$ 1860.8880	0.2262 $\pm$ 0.0405	76650.0000 $\pm$ 13750.9091
Model-free Dijkstra	0.00248 $\pm$ 0.0005	4530.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Asynchronous Value Iteration	0.02155 $\pm$ 0.0005	30177.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Value Iteration	0.02846 $\pm$ 0.0005	41589.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan

Algorithm (Problem 12)	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)
Q-learning ( $\epsilon = 0$ )	0.39384 $\pm$ 0.0030	126594.0000 $\pm$ 0.0000	0.0 %	0.0136 $\pm$ 0.0001	4334.0000 $\pm$ 0.0000	0.1777 $\pm$ 0.0013	57000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 0.25$ )	0.52039 $\pm$ 0.0040	164876.6200 $\pm$ 443.3426	0.0 %	0.0131 $\pm$ 0.0074	4140.6400 $\pm$ 2339.7309	0.2097 $\pm$ 0.0035	66620.0000 $\pm$ 1017.6443
Q-learning ( $\epsilon = 0.5$ )	0.74970 $\pm$ 0.0053	242867.3400 $\pm$ 945.7341	0.0 %	0.0177 $\pm$ 0.0099	5750.2200 $\pm$ 3199.3502	0.2555 $\pm$ 0.0055	83200.0000 $\pm$ 1685.2300
Q-learning ( $\epsilon = 0.75$ )	1.45247 $\pm$ 0.0116	485869.0000 $\pm$ 2962.9134	0.0 %	0.0298 $\pm$ 0.0134	10033.1200 $\pm$ 4511.6419	0.3467 $\pm$ 0.0110	116750.0000 $\pm$ 3488.1944
Q-learning ( $\epsilon = 0.9$ )	1.03468 $\pm$ 0.2863	354415.0400 $\pm$ 97640.2752	100.0 %	0.0595 $\pm$ 0.0288	20555.3000 $\pm$ 9892.0178	0.4977 $\pm$ 0.0240	171190.0000 $\pm$ 7955.7464
Q-learning ( $\epsilon = 1$ )	5.17925 $\pm$ 1.6084	1867336.2200 $\pm$ 579915.6964	94.0 %	0.3157 $\pm$ 0.2704	114299.5000 $\pm$ 97747.7976	4.5908 $\pm$ 1.4548	1654969.3878 $\pm$ 523906.9954
Q-learning ( $\pi$ ) ( $\epsilon = 1$ )	5.26599 $\pm$ 1.0266	1793988.3800 $\pm$ 348867.4422	100.0 %	0.1741 $\pm$ 0.1807	59990.7000 $\pm$ 62199.7318	5.1583 $\pm$ 1.0449	1757340.0000 $\pm$ 355406.9279
Model-free Dijkstra	0.00181 $\pm$ 0.0000	3006.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Asynchronous Value Iteration	0.04413 $\pm$ 0.0007	61435.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Value Iteration	0.06277 $\pm$ 0.0002	91955.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan

Algorithm (Problem 13)	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)
Q-learning ( $\epsilon = 0$ )	0.15450 $\pm$ 0.0019	47576.0000 $\pm$ 0.0000	0.0 %	0.0041 $\pm$ 0.0000	1260.0000 $\pm$ 0.0000	0.0649 $\pm$ 0.0007	20000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 0.25$ )	0.20974 $\pm$ 0.0027	63528.4200 $\pm$ 262.0945	0.0 %	0.0084 $\pm$ 0.0023	2589.2200 $\pm$ 721.6648	0.0759 $\pm$ 0.0014	23110.0000 $\pm$ 371.3489
Q-learning ( $\epsilon = 0.5$ )	0.30891 $\pm$ 0.0039	96395.7400 $\pm$ 547.0078	0.0 %	0.0103 $\pm$ 0.0028	3261.4400 $\pm$ 901.9227	0.0911 $\pm$ 0.0022	28630.0000 $\pm$ 559.5534
Q-learning ( $\epsilon = 0.75$ )	0.66353 $\pm$ 0.0088	214642.2200 $\pm$ 2092.3259	0.0 %	0.0130 $\pm$ 0.0041	4265.2800 $\pm$ 1350.3880	0.1215 $\pm$ 0.0043	39680.0000 $\pm$ 1362.9380
Q-learning ( $\epsilon = 0.9$ )	0.46075 $\pm$ 0.0709	152405.0000 $\pm$ 23204.2928	100.0 %	0.0140 $\pm$ 0.0074	4669.0000 $\pm$ 2496.5186	0.1786 $\pm$ 0.0122	59420.0000 $\pm$ 4108.9658
Q-learning ( $\epsilon = 1$ )	0.72476 $\pm$ 0.1851	251488.8400 $\pm$ 64321.8408	100.0 %	0.0128 $\pm$ 0.0116	4440.9400 $\pm$ 4085.6970	0.4882 $\pm$ 0.1385	169640.0000 $\pm$ 48080.4576
Q-learning ( $\pi$ ) ( $\epsilon = 1$ )	0.82376 $\pm$ 0.2249	278237.9200 $\pm$ 75499.1909	100.0 %	0.0109 $\pm$ 0.0109	3722.1400 $\pm$ 3760.4954	0.5312 $\pm$ 0.1547	179730.0000 $\pm$ 51951.8729
Model-free Dijkstra	0.00212 $\pm$ 0.0005	3874.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Asynchronous Value Iteration	0.01926 $\pm$ 0.0005	27870.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Value Iteration	0.02805 $\pm$ 0.0005	41390.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan

Algorithm (Problem 14)	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)
Q-learning ( $\epsilon = 0$ )	0.11687 $\pm$ 0.0011	36008.0000 $\pm$ 0.0000	0.0 %	0.0053 $\pm$ 0.0000	1615.0000 $\pm$ 0.0000	0.0517 $\pm$ 0.0005	16000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 0.25$ )	0.15673 $\pm$ 0.0016	48378.9200 $\pm$ 210.8762	0.0 %	0.0078 $\pm$ 0.0023	2411.4400 $\pm$ 713.8361	0.0620 $\pm$ 0.0016	19250.0000 $\pm$ 433.0127
Q-learning ( $\epsilon = 0.5$ )	0.22881 $\pm$ 0.0030	72961.0000 $\pm$ 601.7355	0.0 %	0.0098 $\pm$ 0.0026	3142.1200 $\pm$ 846.0950	0.0729 $\pm$ 0.0020	23390.0000 $\pm$ 598.2474
Q-learning ( $\epsilon = 0.75$ )	0.47538 $\pm$ 0.0067	156573.9700 $\pm$ 1738.3132	0.0 %	0.0113 $\pm$ 0.0042	3729.0000 $\pm$ 1393.5140	0.0978 $\pm$ 0.0038	32440.0000 $\pm$ 1210.9500
Q-learning ( $\epsilon = 0.9$ )	0.48236 $\pm$ 0.0984	163299.2600 $\pm$ 32972.2192	100.0 %	0.0095 $\pm$ 0.0062	3224.5600 $\pm$ 2110.4774	0.1341 $\pm$ 0.0069	45590.0000 $\pm$ 2293.8832
Q-learning ( $\epsilon = 1$ )	0.30029 $\pm$ 0.0322	106857.6900 $\pm$ 11388.7065	100.0 %	0.0057 $\pm$ 0.0045	2018.8400 $\pm$ 1595.4904	0.1874 $\pm$ 0.0205	66620.0000 $\pm$ 7206.6358
Q-learning ( $\pi$ ) ( $\epsilon = 1$ )	0.31998 $\pm$ 0.0378	108522.8800 $\pm$ 12668.4291	100.0 %	0.0054 $\pm$ 0.0050	1817.5800 $\pm$ 1713.0152	0.1942 $\pm$ 0.0222	65910.0000 $\pm$ 7482.1053
Model-free Dijkstra	0.00220 $\pm$ 0.0005	3711.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Asynchronous Value Iteration	0.02243 $\pm$ 0.0005	30808.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Value Iteration	0.02449 $\pm$ 0.0001	35112.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan

Algorithm (Problem 15)	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)
Q-learning ( $\epsilon = 0$ )	0.11131 $\pm$ 0.0010	33372.0000 $\pm$ 0.0000	0.0 %	0.0040 $\pm$ 0.0000	1207.0000 $\pm$ 0.0000	0.0599 $\pm$ 0.0005	18000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 0.25$ )	0.14548 $\pm$ 0.0015	43947.8600 $\pm$ 277.4724	0.0 %	0.0071 $\pm$ 0.0016	2168.5400 $\pm$ 506.3488	0.0703 $\pm$ 0.0018	21410.0000 $\pm$ 511.7617
Q-learning ( $\epsilon = 0.5$ )	0.20891 $\pm$ 0.0020	64789.3400 $\pm$ 537.2428	0.0 %	0.0081 $\pm$ 0.0020	2559.7000 $\pm$ 650.5667	0.0855 $\pm$ 0.0029	26800.0000 $\pm$ 894.4272
Q-learning ( $\epsilon = 0.75$ )	0.40536 $\pm$ 0.0057	130578.3200 $\pm$ 1350.2966	0.0 %	0.0085 $\pm$ 0.0036	2766.3800 $\pm$ 1169.2164	0.1189 $\pm$ 0.0062	38750.0000 $\pm$ 1986.8316
Q-learning ( $\epsilon = 0.9$ )	0.95809 $\pm$ 0.0194	315815.5100 $\pm$ 5887.7182	0.0 %	0.0065 $\pm$ 0.0040	2155.7000 $\pm$ 1334.3122	0.1624 $\pm$ 0.0131	54110.0000 $\pm$ 4202.0336
Q-learning ( $\epsilon = 1$ )	0.40327 $\pm$ 0.0676	141574.9600 $\pm$ 23520.4918	100.0 %	0.0037 $\pm$ 0.0030	1279.2800 $\pm$ 1055.8500	0.2170 $\pm$ 0.0238	76250.0000 $\pm$ 8462.1215
Q-learning ( $\pi$ ) ( $\epsilon = 1$ )	0.39702 $\pm$ 0.0784	135803.2500 $\pm$ 26665.6822	100.0 %	0.0029 $\pm$ 0.0027	994.5000 $\pm$ 942.6673	0.2230 $\pm$ 0.0282	76310.0000 $\pm$ 9585.0874
Model-free Dijkstra	0.00253 $\pm$ 0.0007	4494.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Asynchronous Value Iteration	0.01561 $\pm$ 0.0002	22175.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Value Iteration	0.02710 $\pm$ 0.0005	39619.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan

Algorithm (Problem 16)	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)
Q-learning ( $\epsilon = 0$ )	0.00285 $\pm$ 0.0000	1000.0000 $\pm$ 0.0000	100.0 %	0.0000 $\pm$ 0.0000	14.0000 $\pm$ 0.0000	0.0028 $\pm$ 0.0000	1000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 0.25$ )	0.00289 $\pm$ 0.0000	1002.1600 $\pm$ 2.1294	100.0 %	0.0001 $\pm$ 0.0000	16.0000 $\pm$ 3.3941	0.0029 $\pm$ 0.0000	1000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 0.5$ )	0.00283 $\pm$ 0.0000	1004.0000 $\pm$ 3.7094	100.0 %	0.0001 $\pm$ 0.0000	18.3000 $\pm$ 7.0050	0.0028 $\pm$ 0.0000	1000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 0.75$ )	0.00278 $\pm$ 0.0000	1006.4000 $\pm$ 6.9685	100.0 %	0.0001 $\pm$ 0.0000	20.9800 $\pm$ 10.2781	0.0028 $\pm$ 0.0000	1000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 0.9$ )	0.00276 $\pm$ 0.0000	1013.2800 $\pm$ 14.3095	100.0 %	0.0001 $\pm$ 0.0001	26.3400 $\pm$ 20.9496	0.0027 $\pm$ 0.0000	1000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 1$ )	0.00266 $\pm$ 0.0001	1019.6000 $\pm$ 22.4410	100.0 %	0.0001 $\pm$ 0.0001	23.6000 $\pm$ 19.2603	0.0026 $\pm$ 0.0000	1000.0000 $\pm$ 0.0000
Q-learning ( $\pi$ ) ( $\epsilon = 1$ )	0.00307 $\pm$ 0.0001	1023.6400 $\pm$ 24.1386	100.0 %	0.0001 $\pm$ 0.0001	21.6400 $\pm$ 22.2484	0.0030 $\pm$ 0.0000	1000.0

**Table 4.** Algorithm performance on Problem 1, run on a desktop computer with an i5-12400F CPU (2.50GHz). Includes Q-learning with  $\epsilon$ -decay.

Algorithm (Problem 1)	Runtime (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)
Q-learning ( $\epsilon = 0$ )	0.12396 $\pm$ 0.0077	40894.0000 $\pm$ 0.0000	0.0 %	0.0033 $\pm$ 0.0009	1127.0000 $\pm$ 0.0000	0.0605 $\pm$ 0.0055	20000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 0.25$ )	0.16006 $\pm$ 0.0090	53322.1200 $\pm$ 224.6660	0.0 %	0.0065 $\pm$ 0.0026	2143.6200 $\pm$ 770.4978	0.0707 $\pm$ 0.0074	23530.0000 $\pm$ 499.0992
Q-learning ( $\epsilon = 0.5$ )	0.23343 $\pm$ 0.0113	79207.6000 $\pm$ 547.0679	0.0 %	0.0093 $\pm$ 0.0037	3057.7400 $\pm$ 979.0199	0.0874 $\pm$ 0.0066	29590.0000 $\pm$ 722.4265
Q-learning ( $\epsilon = 0.75$ )	0.49495 $\pm$ 0.0278	172365.5800 $\pm$ 7977.4984	10.0 %	0.0106 $\pm$ 0.0043	3679.7600 $\pm$ 1235.4283	0.1206 $\pm$ 0.0092	42170.0000 $\pm$ 1697.3803
Q-learning ( $\epsilon = 0.9$ )	0.25825 $\pm$ 0.0410	92243.3600 $\pm$ 14580.8509	100.0 %	0.0092 $\pm$ 0.0053	3264.8600 $\pm$ 1844.4111	0.1650 $\pm$ 0.0133	58850.0000 $\pm$ 3561.9517
Q-learning ( $\epsilon = 1$ )	0.30633 $\pm$ 0.0452	114753.3800 $\pm$ 16236.2484	100.0 %	0.0039 $\pm$ 0.0038	1413.7000 $\pm$ 1342.0072	0.2580 $\pm$ 0.0400	96400.0000 $\pm$ 14849.2424
Q-learning ( $\pi$ ) ( $\epsilon = 1$ )	0.30954 $\pm$ 0.0371	114637.3700 $\pm$ 13326.4214	100.0 %	0.0046 $\pm$ 0.0039	1643.4800 $\pm$ 1398.7443	0.2630 $\pm$ 0.0358	97310.0000 $\pm$ 12780.2152
Q-learning ( $\epsilon$ -decay)	0.29742 $\pm$ 0.1000	104886.2900 $\pm$ 25156.3045	93.0 %	0.0053 $\pm$ 0.0071	1847.1000 $\pm$ 1657.8496	0.2229 $\pm$ 0.0522	78710.0000 $\pm$ 6521.1885
Model-free Dijkstra	0.00202 $\pm$ 0.0007	4015.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Asynchronous Value Iteration	0.01113 $\pm$ 0.0026	23308.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Value Iteration	0.01554 $\pm$ 0.0026	35078.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan

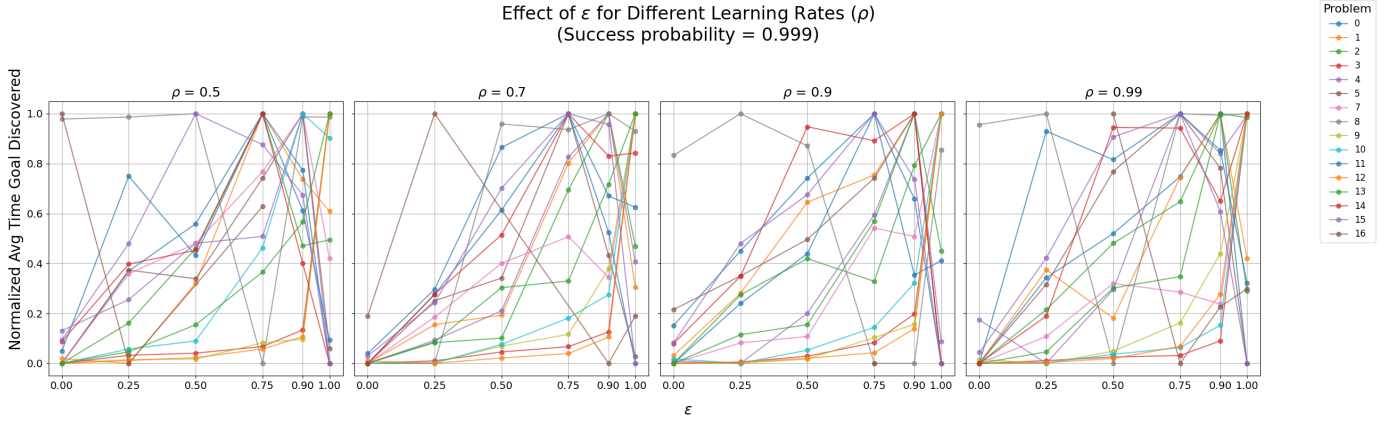
**Table 5.** Algorithm performance on Problem 10, run on a desktop computer with an i5-12400F CPU (2.50GHz). Includes Q-learning with  $\epsilon$ -decay.

Algorithm (Problem 10)	Runtime (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)
Q-learning ( $\epsilon = 0$ )	0.29829 $\pm$ 0.0188	101214.0000 $\pm$ 0.0000	0.0 %	0.0061 $\pm$ 0.0013	2015.0000 $\pm$ 0.0000	0.1530 $\pm$ 0.0123	52000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 0.25$ )	0.39699 $\pm$ 0.0211	133139.2000 $\pm$ 427.4064	0.0 %	0.0089 $\pm$ 0.0049	3046.7600 $\pm$ 1631.0804	0.1848 $\pm$ 0.0117	61960.0000 $\pm$ 1085.5413
Q-learning ( $\epsilon = 0.5$ )	0.57380 $\pm$ 0.0171	200169.4200 $\pm$ 1030.5368	0.0 %	0.0141 $\pm$ 0.0065	4878.0600 $\pm$ 2268.3855	0.2258 $\pm$ 0.0121	78600.0000 $\pm$ 1870.8287
Q-learning ( $\epsilon = 0.75$ )	0.92641 $\pm$ 0.1920	333304.1300 $\pm$ 68721.9494	83.0 %	0.0210 $\pm$ 0.0076	7685.2200 $\pm$ 2637.8776	0.3114 $\pm$ 0.0163	112850.0000 $\pm$ 3612.1323
Q-learning ( $\epsilon = 0.9$ )	0.54554 $\pm$ 0.0515	203946.4000 $\pm$ 18464.7784	100.0 %	0.0333 $\pm$ 0.0180	12237.4000 $\pm$ 6515.6084	0.4278 $\pm$ 0.0261	160460.0000 $\pm$ 7338.1469
Q-learning ( $\epsilon = 1$ )	1.69006 $\pm$ 0.5458	658601.9700 $\pm$ 211027.5533	100.0 %	0.0649 $\pm$ 0.0751	25902.0000 $\pm$ 29790.2904	1.1196 $\pm$ 0.2805	437430.0000 $\pm$ 108380.5568
Q-learning ( $\pi$ ) ( $\epsilon = 1$ )	1.78199 $\pm$ 0.4455	685022.7200 $\pm$ 169670.6191	100.0 %	0.0472 $\pm$ 0.0505	18366.3400 $\pm$ 19499.6667	1.1819 $\pm$ 0.2485	455760.0000 $\pm$ 94888.6843
Q-learning ( $\epsilon$ -decay)	0.65391 $\pm$ 0.1155	260098.3800 $\pm$ 39405.3940	98.0 %	0.0624 $\pm$ 0.0455	25292.6600 $\pm$ 18330.5052	0.5338 $\pm$ 0.0508	213360.0000 $\pm$ 13813.4138
Model-free Dijkstra	0.00229 $\pm$ 0.0006	4423.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Asynchronous Value Iteration	0.03125 $\pm$ 0.0032	69570.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Value Iteration	0.03501 $\pm$ 0.0028	82594.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan

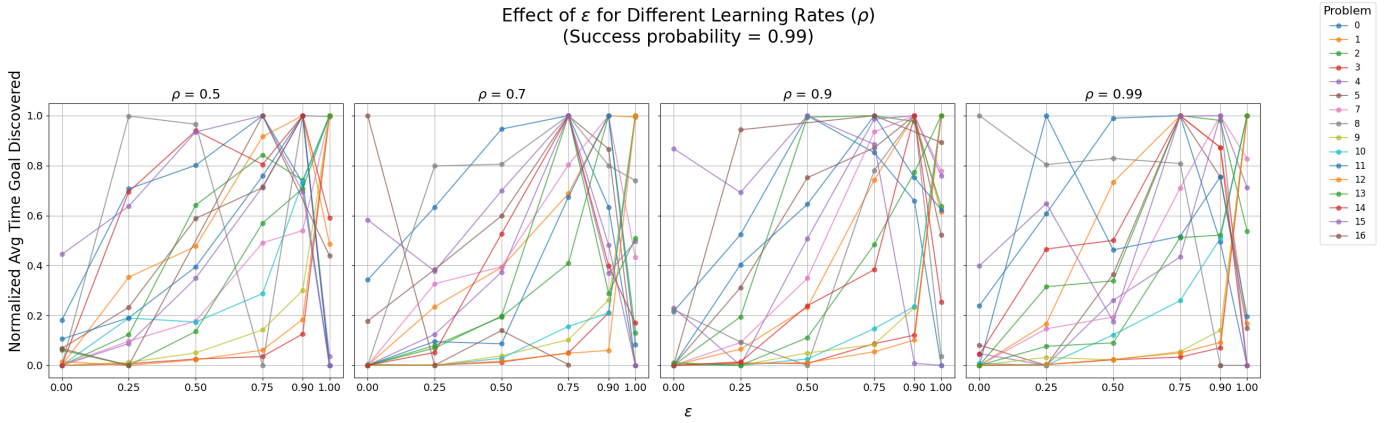
**Table 6.** Algorithm performance on Problem 11, run on a desktop computer with an i5-12400F CPU (2.50GHz). Includes Q-learning with  $\epsilon$ -decay.

Algorithm (Problem 11)	Runtime (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)
Q-learning ( $\epsilon = 0$ )	0.10471 $\pm$ 0.0075	35316.0000 $\pm$ 0.0000	0.0 %	0.0047 $\pm$ 0.0011	1569.0000 $\pm$ 0.0000	0.0501 $\pm$ 0.0039	17000.0000 $\pm$ 0.0000
Q-learning ( $\epsilon = 0.25$ )	0.14444 $\pm$ 0.0083	48838.7600 $\pm$ 311.0242	0.0 %	0.0061 $\pm$ 0.0019	2081.9200 $\pm$ 537.3621	0.0601 $\pm$ 0.0056	20320.0000 $\pm$ 487.4423
Q-learning ( $\epsilon = 0.5$ )	0.21523 $\pm$ 0.0092	74975.5100 $\pm$ 580.9994	0.0 %	0.0073 $\pm$ 0.0029	2529.5200 $\pm$ 889.7264	0.0712 $\pm$ 0.0061	24870.0000 $\pm$ 770.1299
Q-learning ( $\epsilon = 0.75$ )	0.46522 $\pm$ 0.0216	166059.1000 $\pm$ 2033.6388	0.0 %	0.0095 $\pm$ 0.0046	3428.7800 $\pm$ 1434.6559	0.0961 $\pm$ 0.0076	34520.0000 $\pm$ 1590.4716
Q-learning ( $\epsilon = 0.9$ )	0.73656 $\pm$ 0.1717	270569.9300 $\pm$ 60522.7662	98.0 %	0.0085 $\pm$ 0.0055	3121.8000 $\pm$ 2038.0453	0.1253 $\pm$ 0.0110	46730.0000 $\pm$ 2918.4071
Q-learning ( $\epsilon = 1$ )	0.36355 $\pm$ 0.0525	141925.4500 $\pm$ 17284.5134	100.0 %	0.0055 $\pm$ 0.0051	2234.1800 $\pm$ 2136.7621	0.1997 $\pm$ 0.0398	77680.0000 $\pm$ 14925.7362
Q-learning ( $\pi$ ) ( $\epsilon = 1$ )	0.37555 $\pm$ 0.0586	144673.6700 $\pm$ 21823.7557	100.0 %	0.0051 $\pm$ 0.0051	1936.9800 $\pm$ 1860.8880	0.1989 $\pm$ 0.0351	76650.0000 $\pm$ 13750.9091
Q-learning ( $\epsilon$ -decay)	0.47468 $\pm$ 0.0362	185298.8500 $\pm$ 6114.1593	0.0 %	0.0046 $\pm$ 0.0071	1880.0600 $\pm$ 1531.7343	0.1569 $\pm$ 0.0231	63160.0000 $\pm$ 6067.4871
Model-free Dijkstra	0.00258 $\pm$ 0.0010	4530.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Asynchronous Value Iteration	0.01403 $\pm$ 0.0014	30177.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan
Model-free Value Iteration	0.01785 $\pm$ 0.0017	41589.0000 $\pm$ 0.0000	100.0 %	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan	nan $\pm$ nan

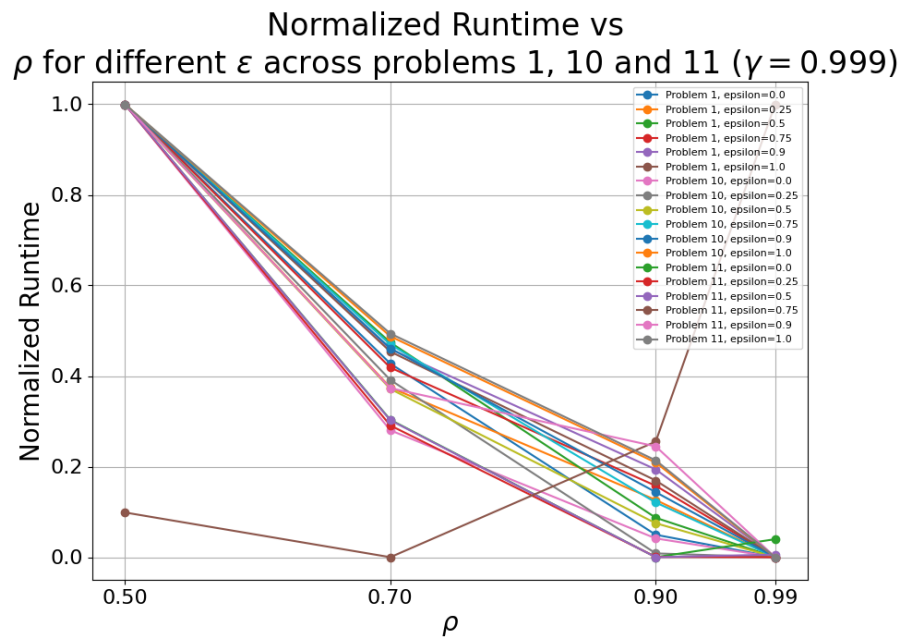
C Figures for stochastic experiments



**Fig. 6.** The effect of decreasing greediness in a policy when applied on different problems with  $\gamma = 0.999$  and varying learning rate  $\rho$ .



**Fig. 7.** The effect of decreasing greediness in a policy when applied on different problems with  $\gamma = 0.99$  and varying learning rate  $\rho$ .



**Fig. 8.** As  $\rho$  increases, the run time for Q-learning decreases when  $\gamma = 0.999$  for all  $\epsilon$  with the exception of Problem 11 with  $\epsilon = 0.75$ .

## D Stochastic experiments

Bellow we have the tables for Problems 1,10 and 11 with  $\gamma \in \{0.99, 0.99\}$ . These algorithms were run  $N = 100$  times. The table shows similar metrics as in the deterministic case, but we also list how often the optimal cost-to-go value for the initial state is achieved, and the shortest and longest paths extracted from the computed values. Afterwards, the tables for Problems 1,10 and 11 with  $\gamma \in \{0.9, 0.7, 0.6\}$  are shown. These were run  $N = 10$  times since they were considerably slower due to the stochasticity. We also used a desktop computer with an i5-12400F CPU (2.50GHz) for this batch of experiments and the ones we describe below. We show results using  $\epsilon$ -decay in Tables 22, 23 and 24. Additional experiments with low  $\gamma$  for problems 8 and 16 are also included. Finally, the tables for the rest of the problems ran with  $\gamma \in \{0.99, 0.999\}$  are shown.

**Table 7.** Performance of DP methods and Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.99$  (Problem 1).

Algorithm (Problem 1) ( $\gamma = 0.99$ )	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	0.44775 $\pm$ 0.0094	110018.9500 $\pm$ 120.6576	0.0%	0.0041 $\pm$ 0.0013	1019.9000 $\pm$ 300.0232	0.1815 $\pm$ 0.0082	4680.0000 $\pm$ 1907.2493	100.0%	28/34
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	0.42652 $\pm$ 0.0104	103655.8100 $\pm$ 848.3384	0.0%	0.0040 $\pm$ 0.0012	968.4300 $\pm$ 267.7274	0.1359 $\pm$ 0.0100	33040.0000 $\pm$ 2262.3881	100.0%	28/1456
Q-learning ( $\rho = 0.9, \epsilon = 0$ )	0.41722 $\pm$ 0.0076	102035.4600 $\pm$ 1339.1680	0.0%	0.0042 $\pm$ 0.0013	1014.5900 $\pm$ 292.7621	0.1306 $\pm$ 0.0429	31970.0000 $\pm$ 10627.7514	100.0%	28/920
Q-learning ( $\rho = 0.99, \epsilon = 0$ )	0.42034 $\pm$ 0.0090	102411.8500 $\pm$ 1536.8528	0.0%	0.0042 $\pm$ 0.0012	1026.1100 $\pm$ 282.5215	0.2078 $\pm$ 0.0750	30697.6744 $\pm$ 18153.7869	86.0%	28/2205
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	0.56925 $\pm$ 0.0061	141115.7100 $\pm$ 320.6419	0.0%	0.0050 $\pm$ 0.0015	1472.4300 $\pm$ 377.8420	0.3409 $\pm$ 0.0901	84917.6471 $\pm$ 22314.9708	83.0%	28/82
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	0.54557 $\pm$ 0.0069	134211.4300 $\pm$ 390.0833	0.0%	0.0050 $\pm$ 0.0014	1222.1600 $\pm$ 355.0452	0.3071 $\pm$ 0.1184	75800.0000 $\pm$ 28881.8282	50.0%	28/31
Q-learning ( $\rho = 0.9, \epsilon = 0.25$ )	0.52808 $\pm$ 0.0066	130873.3000 $\pm$ 466.7216	0.0%	0.0052 $\pm$ 0.0015	1279.0000 $\pm$ 358.8929	0.2649 $\pm$ 0.1150	65640.0000 $\pm$ 28480.0000	25.0%	28/31
Q-learning ( $\rho = 0.99, \epsilon = 0.25$ )	0.52894 $\pm$ 0.0130	130261.6100 $\pm$ 474.9027	0.0%	0.0054 $\pm$ 0.0016	1322.2200 $\pm$ 398.5655	0.3375 $\pm$ 0.0900	83000.0000 $\pm$ 0.0000	1.0%	28/31
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	0.77585 $\pm$ 0.0169	197307.0000 $\pm$ 857.2310	0.0%	0.0053 $\pm$ 0.0017	1622.1800 $\pm$ 433.5110	nan $\pm$ nan	nan $\pm$ nan	0.0%	28/1474
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	0.82059 $\pm$ 0.0169	206637.7200 $\pm$ 779.9925	0.0%	0.0073 $\pm$ 0.0018	1864.0100 $\pm$ 450.2174	0.5143 $\pm$ 0.1522	129593.5700 $\pm$ 37137.2954	32.0%	28/31
Q-learning ( $\rho = 0.9, \epsilon = 0.5$ )	0.79430 $\pm$ 0.0118	200809.2200 $\pm$ 926.3823	0.0%	0.0067 $\pm$ 0.0017	1694.9400 $\pm$ 416.3421	0.4310 $\pm$ 0.1809	109380.9524 $\pm$ 45658.6467	21.0%	28/31
Q-learning ( $\rho = 0.99, \epsilon = 0.5$ )	0.78556 $\pm$ 0.0137	198195.8800 $\pm$ 936.3099	0.0%	0.0065 $\pm$ 0.0020	1626.9000 $\pm$ 505.6041	0.3918 $\pm$ 0.1863	100000.0000 $\pm$ 47588.5140	3.0%	28/31
Q-learning ( $\rho = 0.99, \epsilon = 0.75$ )	1.62392 $\pm$ 0.0210	422192.3000 $\pm$ 3032.9179	0.0%	0.0084 $\pm$ 0.0029	2187.4200 $\pm$ 769.7782	nan $\pm$ nan	nan $\pm$ nan	0.0%	28/31
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	1.57260 $\pm$ 0.0235	413107.3800 $\pm$ 2630.4240	0.0%	0.0093 $\pm$ 0.0033	2458.8300 $\pm$ 861.9200	0.9108 $\pm$ 0.3088	240235.2941 $\pm$ 80468.8682	34.0%	28/31
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	1.56681 $\pm$ 0.0272	410248.2100 $\pm$ 3179.0668	0.0%	0.0087 $\pm$ 0.0035	2291.0500 $\pm$ 915.9052	0.5933 $\pm$ 0.3101	155533.3333 $\pm$ 81770.3016	15.0%	28/31
Q-learning ( $\rho = 0.9, \epsilon = 0.75$ )	1.56187 $\pm$ 0.0176	411440.7600 $\pm$ 3119.8781	0.0%	0.0086 $\pm$ 0.0027	2261.1900 $\pm$ 731.0403	1.4754 $\pm$ 0.0000	391000.0000 $\pm$ 0.0000	1.0%	28/30
Q-learning ( $\rho = 0.99, \epsilon = 0.75$ )	3.84191 $\pm$ 0.0587	1027729.5600 $\pm$ 11088.5495	0.0%	0.0085 $\pm$ 0.0048	2290.7600 $\pm$ 1295.1463	2.0426 $\pm$ 1.0165	547480.0000 $\pm$ 272473.1356	25.0%	28/31
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	3.82287 $\pm$ 0.1084	1019592.4000 $\pm$ 10171.1316	0.0%	0.0099 $\pm$ 0.0048	2652.3500 $\pm$ 1285.4855	2.0493 $\pm$ 0.9742	549081.6327 $\pm$ 260621.2231	49.0%	28/31
Q-learning ( $\rho = 0.7, \epsilon = 0.9$ )	3.84191 $\pm$ 0.0587	1027729.5600 $\pm$ 11088.5495	0.0%	0.0085 $\pm$ 0.0048	2290.7600 $\pm$ 1295.1463	2.0426 $\pm$ 1.0165	547480.0000 $\pm$ 272473.1356	25.0%	28/31
Q-learning ( $\rho = 0.9, \epsilon = 0.9$ )	4.00248 $\pm$ 0.0630	1081800.3000 $\pm$ 12373.4476	0.0%	0.0091 $\pm$ 0.0056	2471.7700 $\pm$ 1518.8372	2.8513 $\pm$ 0.9324	768333.3333 $\pm$ 246698.6466	3.0%	28/33
Q-learning ( $\rho = 0.99, \epsilon = 0.9$ )	4.85966 $\pm$ 0.0606	1180762.1000 $\pm$ 13875.0949	0.0%	0.0092 $\pm$ 0.0045	2094.2500 $\pm$ 1240.8426	nan $\pm$ nan	nan $\pm$ nan	0.0%	28/532
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	15.46274 $\pm$ 0.1914	4324185.7000 $\pm$ 48051.5290	0.0%	0.0053 $\pm$ 0.0040	1481.3700 $\pm$ 1126.3129	5.2842 $\pm$ 3.5571	1478922.2222 $\pm$ 997581.0992	90.0%	28/33
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	15.38096 $\pm$ 0.2003	4319096.5300 $\pm$ 52853.4397	0.0%	0.0057 $\pm$ 0.0051	1588.3300 $\pm$ 1441.8374	5.6281 $\pm$ 4.1104	1579675.0000 $\pm$ 1155420.2458	80.0%	28/31
Q-learning ( $\rho = 0.9, \epsilon = 1$ )	15.39582 $\pm$ 0.2069	4319667.6700 $\pm$ 54011.2494	0.0%	0.0059 $\pm$ 0.0048	1638.9100 $\pm$ 1327.6757	7.1078 $\pm$ 4.6024	1999958.3333 $\pm$ 1293126.1372	24.0%	28/3346
Q-learning ( $\rho = 0.99, \epsilon = 1$ )	15.39603 $\pm$ 0.1846	4314001.5500 $\pm$ 48947.1783	0.0%	0.0060 $\pm$ 0.0048	1661.0400 $\pm$ 1336.9004	nan $\pm$ nan	nan $\pm$ nan	0.0%	28/42
Async Value Iteration	0.15656 $\pm$ 0.0014	137472.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	28/33
Value Iteration	0.22934 $\pm$ 0.0020	194752.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	28/31

**Table 8.** Performance of DP methods and Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.999$  (Problem 1).

Algorithm (Problem 1) ( $\gamma = 0.999$ )	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	0.45780 $\pm$ 0.0072	108128.1300 $\pm$ 66.4493	0.0%	0.0046 $\pm$ 0.0004	1072.3100 $\pm$ 94.0911	0.1763 $\pm$ 0.0038	41650.0000 $\pm$ 476.9696	100.0%	28/7264
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	0.42838 $\pm$ 0.0087	107074.8600 $\pm$ 108.2371	0.0%	0.0045 $\pm$ 0.0005	1053.1400 $\pm$ 1124.3148	0.1273 $\pm$ 0.0027	29950.0000 $\pm$ 217.9449	100.0%	28/1456
Q-learning ( $\rho = 0.9, \epsilon = 0$ )	0.40930 $\pm$ 0.0046	96842.1900 $\pm$ 116.4245	0.0%	0.0044 $\pm$ 0.0006	1041.3000 $\pm$ 148.1313	0.0973 $\pm$ 0.0012	23020.0000 $\pm$ 140.0000	100.0%	28/30
Q-learning ( $\rho = 0.99, \epsilon = 0$ )	0.40677 $\pm$ 0.0041	96121.0500 $\pm$ 195.6695	0.0%	0.0044 $\pm$ 0.0006	1044.4500 $\pm$ 130.2082	0.0888 $\pm$ 0.0008	21000.0000 $\pm$ 0.0000	100.0%	28/32
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	0.57845 $\pm$ 0.0051	138790.8000 $\pm$ 386.8301	0.0%	0.0061 $\pm$ 0.0017	1468.7500 $\pm$ 399.6031	0.1898 $\pm$ 0.0019	45920.0000 $\pm$ 305.9412	100.0%	28/30
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	0.54707 $\pm$ 0.0058	131788.1800 $\pm$ 402.8578	0.0%	0.0050 $\pm$ 0.0014	1217.3300 $\pm$ 349.6271	0.1381 $\pm$ 0.0026	33490.0000 $\pm$ 5195.1900	100.0%	28/30
Q-learning ( $\rho = 0.9, \epsilon = 0.25$ )	0.53586 $\pm$ 0.0064	128248.1100 $\pm$ 438.3435	0.0%	0.0051 $\pm$ 0.0015	1239.5400 $\pm$ 340.2047	0.1100 $\pm$ 0.0024	26560.0000 $\pm$ 496.3869	100.0%	28/29
Q-learning ( $\rho = 0.99, \epsilon = 0.25$ )	0.52970 $\pm$ 0.0064	127530.0700 $\pm$ 370.5002	0.0%	0.0053 $\pm$ 0.0016	1278.4100 $\pm$ 378.5350	0.1006 $\pm$ 0.0024	24320.0000 $\pm$ 466.4762	100.0%	28/29
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	0.81914 $\pm$ 0.0101	203280.6600 $\pm$ 1146.9714	0.0%	0.0073 $\pm$ 0.0020	1814.4300 $\pm$ 480.8558	0.2137 $\pm$ 0.0035	53480.0000 $\pm$ 556.4171	100.0%	28/30
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	0.78860 $\pm$ 0.0079	197246.0000 $\pm$ 978.0357	0.0%	0.0068 $\pm$ 0.0018	1702.0100 $\pm$ 435.4708	0.1581 $\pm$ 0.0025	38000.0000 $\pm$ 489.8979	100.0%	28/31
Q-learning ( $\rho = 0.9, \epsilon = 0.5$ )	0.76821 $\pm$ 0.0160	196367.9900 $\pm$ 1204.9415	0.0%	0.0064 $\pm$ 0.0020	1632.6200 $\pm$ 507.3923	0.1254 $\pm$ 0.0032	32350.0000 $\pm$ 606.2178	100.0%	28/30
Q-learning ( $\rho = 0.99, \epsilon = 0.5$ )	0.76335 $\pm$ 0.0084	197826.2100 $\pm$ 875.8254	0.0%	0.0063 $\pm$ 0.0016	1634.3100 $\pm$ 421.2690	0.1144 $\pm$ 0.0024	29800.0000 $\pm$ 529.1503	100.0%	28/30
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	1.20039 $\pm$ 0.2329	321255.2700 $\pm$ 62103.2471	88.0%	0.0088 $\pm$ 0.0033	2362.1600 $\pm$ 872.5669	0.2545 $\pm$ 0.0053	68620.0000 $\pm$ 1172.8598	100.0%	28/29
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	0.80764 $\pm$ 0.1909	215408.4700 $\pm$ 56702.6464	100.0%	0.0083 $\pm$ 0.0031	2217.5700 $\pm$ 840.3794	0.1969 $\pm$ 0.0063	62760.0000 $\pm$ 1304.7605	100.0%	28/29
Q-learning ( $\rho = 0.9, \epsilon = 0.75$ )	0.63141 $\pm$ 0.1970	169741.5500 $\pm$ 52725.1623	100.0%	0.0091 $\pm$ 0.0037	2441.8600 $\pm$ 986.8430	0.1640 $\pm$ 0.0048	44210.0000 $\pm$ 1194.1105	100.0%	28/29
Q-learning ( $\rho = 0.99, \epsilon = 0.75$ )	0.52489 $\pm$ 0.1930	140000.4800 $\pm$ 50657.7332	100.0%	0.0090 $\pm$ 0.0034	2408.0900 $\pm$ 894.9593	0.1516 $\pm$ 0.0045	40480.0000 $\pm$ 1081.4805	100.0%	28/30
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	0.51857 $\pm$ 0.0569	140998.0100 $\pm$ 15374.5296	100.0%	0.0093 $\pm$ 0.0056	2524.0900 $\pm$ 1527.4172	0.3294 $\pm$ 0.0125	89500.0000 $\pm$ 3185.9065	100.0%	28/30
Q-learning ( $\rho = 0.7, \epsilon = 0.9$ )	0.38885 $\pm$ 0.0494	109196.5900 $\pm$ 12728.0244	100.0%	0.0095 $\pm$ 0.0053	2586.7900 $\pm$ 1449.9621	0.2606 $\pm$ 0.0111	71060.0000 $\pm$ 3012.7064	100.0%	28/30
Q-learning ( $\rho = 0.9, \epsilon = 0.9$ )	0.32507 $\pm$ 0.0361	89175.6700 $\pm$ 9912.0126	100.0%	0.0084 $\pm$ 0.0049	2299.5000 $\pm$ 1333.3552	0.2203 $\pm$ 0.0109	60350.0000 $\pm$ 2878.8018	100.0%	28/29
Q-learning ( $\rho = 0.99, \epsilon = 0.9$ )	0.27856 $\pm$ 0.0271	75963.0100 $\pm$ 7419.7872	100.0%	0.0089 $\pm$ 0.0046	2421.2300 $\pm$ 1268.8343	0.2044 $\pm$ 0.0109	55670.0000 $\pm$ 2915.6646	100.0%	28/29
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	0.68562 $\pm$ 0.0786	190240.5100 $\pm$ 21205.5439	100.0%	0.0083 $\pm$ 0.0062	1735.3900 $\pm$ 1720.3775	0.6086 $\pm$ 0.0661	168960.0000 $\pm$ 18121.2141	100.0%	28/29
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	0.53277 $\pm$ 0.0642	148527.9800 $\pm$ 17641.4631	100.0%	0.0052 $\pm$ 0.0047	1466.7500 $\pm$ 1296.5192	0.4704 $\pm$ 0.0622	131110.0000 $\pm$ 17159.1929	100.0%	28/30
Q-learning ( $\rho = 0.9, \epsilon = 1$ )	0.45312 $\pm$ 0.0586	126285.1000 $\pm$ 16381.5359	100.0%	0.0064 $\pm$ 0.0048	1770.8200 $\pm$ 1325.8271	0.3915 $\pm$ 0.0569	109050.0000 $\pm$ 15879.7827	100.0%	28/30

**Table 9.** Performance of DP methods and Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.99$  (Problem 10).

Algorithm (Problem 10) ( $\gamma = 0.99$ )	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	1.09601 $\pm$ 0.0226	270673.0900 $\pm$ 1030.1652	0.0%	0.0098 $\pm$ 0.0041	2379.3400 $\pm$ 996.5445	4.524 $\pm$ 0.0133	11080.0000 $\pm$ 2561.3278	100.0%	64/322
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	1.03734 $\pm$ 0.0188	256745.5500 $\pm$ 3275.4224	0.0%	0.0099 $\pm$ 0.0039	2416.5100 $\pm$ 938.7294	0.3480 $\pm$ 0.1000	85742.2680 $\pm$ 25179.4706	97.0%	64/1614
Q-learning ( $\rho = 0.9, \epsilon = 0$ )	1.05040 $\pm$ 0.0185	259065.8600 $\pm$ 3638.1183	0.0%	0.0092 $\pm$ 0.0033	2244.6800 $\pm$ 804.8872	0.3135 $\pm$ 0.1367	77466.6667 $\pm$ 33639.9894	60.0%	66/45219
Q-learning ( $\rho = 0.99, \epsilon = 0$ )	1.07131 $\pm$ 0.0178	262565.8200 $\pm$ 3765.1413	0.0%	0.0093 $\pm$ 0.0036	2259.5400 $\pm$ 896.9925	0.2651 $\pm$ 0.0000	65000.0000 $\pm$ 0.0000	1.0%	66/2596
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	1.43981 $\pm$ 0.0548	354087.5500 $\pm$ 622.1560	0.0%	0.0140 $\pm$ 0.0077	3431.0800 $\pm$ 1861.5218	0.6260 $\pm$ 0.1351	154170.0000 $\pm$ 32728.2920	100.0%	64/70
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	1.34549 $\pm$ 0.0105	334650.5700 $\pm$ 1071.9651	0.0%	0.0105 $\pm$ 0.0067	2620.3600 $\pm$ 1668.5465	0.4836 $\pm$ 0.1409	120480.0000 $\pm$ 35038.4018	100.0%	64/112
Q-learning ( $\rho = 0.9, \epsilon = 0.25$ )	1.34040 $\pm$ 0.0127	328920.8100 $\pm$ 1294.5546	0.0%	0.0118 $\pm$ 0.0073	2921.1600 $\pm$ 1811.3192	0.3797 $\pm$ 0.1185	93410.0000 $\pm$ 27494.7613	100.0%	64/7336
Q-learning ( $\rho = 0.99, \epsilon = 0.25$ )	1.31547 $\pm$ 0.0130	328258.1500 $\pm$ 1489.9179	0.0%	0.0098 $\pm$ 0.0060	2448.4600 $\pm$ 1519.3635	0.5097 $\pm$ 0.2413	127547.3684 $\pm$ 60110.9095	95.0%	64/1706
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	2.08316 $\pm$ 0.0157	531713.7700 $\pm$ 1337.2434	0.0%	0.0181 $\pm$ 0.0094	4630.6000 $\pm$ 2360.6759	0.7859 $\pm$ 0.2102	203550.0000 $\pm$ 53193.3831	100.0%	64/68
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	2.01877 $\pm$ 0.0167	514338.0800 $\pm$ 1448.9041	0.0%	0.0148 $\pm$ 0.0080	3775.6000 $\pm$ 2037.4858	0.6919 $\pm$ 0.2492	176790.0000 $\pm$ 63480.7522	100.0%	64/79
Q-learning ( $\rho = 0.9, \epsilon = 0.5$ )	2.00751 $\pm$ 0.0192	508357.3700 $\pm$ 1698.5668	0.0%	0.0158 $\pm$ 0.0096	4033.5500 $\pm$ 2437.1517	0.5771 $\pm$ 0.2580	146590.0000 $\pm$ 64556.9663	100.0%	64/795
Q-learning ( $\rho = 0.99, \epsilon = 0.5$ )	1.59043 $\pm$ 0.0217	508076.7500 $\pm$ 1609.1026	0.0%	0.0153 $\pm$ 0.0081	3904.9100 $\pm$ 2061.3847	0.6288 $\pm$ 0.3471	160690.0000 $\pm$ 64347.5133	99.0%	64/3989
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	4.16759 $\pm$ 0.0386	1099049.5400 $\pm$ 5207.7051	0.0%	0.0253 $\pm$ 0.0135	6721.1800 $\pm$ 3592.8100	1.0769 $\pm$ 0.3866	285020.0000 $\pm$ 104270.5912	100.0%	64/68
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	4.12914 $\pm$ 0.0413	1091002.3000 $\pm$ 4279.7067	0.0%	0.0226 $\pm$ 0.0131	6016.9800 $\pm$ 3466.9508	0.9614 $\pm$ 0.4025	254810.0000 $\pm$ 105862.1457	100.0%	64/68
Q-learning ( $\rho = 0.9, \epsilon = 0.75$ )	4.12844 $\pm$ 0.0431	1091267.5000 $\pm$ 4094.5104	0.0%	0.0232 $\pm$ 0.0132	6188.1300 $\pm$ 3511.1027	0.8190 $\pm$ 0.4050	217540.0000 $\pm$ 106509.2879	100.0%	64/654
Q-learning ( $\rho = 0.99, \epsilon = 0.75$ )	4.16265 $\pm$ 0.0467	1093979.6700 $\pm$ 4853.2607	0.0%	0.0225 $\pm$ 0.0137	5966.1100 $\pm$ 3618.1057	1.0192 $\pm$ 0.7982	279090.0000 $\pm$ 210896.2640	99.0%	64/1395
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	10.57651 $\pm$ 0.1140	2868949.5500 $\pm$ 18988.1324	0.0%	0.0421 $\pm$ 0.0250	11472.2500 $\pm$ 6795.5047	1.8967 $\pm$ 1.0263	514400.0000 $\pm$ 278406.6450	100.0%	64/69
Q-learning ( $\rho = 0.7, \epsilon = 0.9$ )	10.54571 $\pm$ 0.1132	2879704.9100 $\pm$ 18072.6930	0.0%	0.0376 $\pm$ 0.0232	10316.3800 $\pm$ 6390.1483	1.7248 $\pm$ 0.8627	471640.0000 $\pm$ 234910.3455	100.0%	64/323
Q-learning ( $\rho = 0.9, \epsilon = 0.9$ )	10.65997 $\pm$ 0.1173	2907657.8700 $\pm$ 20629.3163	0.0%	0.0408 $\pm$ 0.0238	11182.0800 $\pm$ 6486.2508	1.5113 $\pm$ 0.9102	413190.0000 $\pm$ 248451.4719	100.0%	64/1372
Q-learning ( $\rho = 0.99, \epsilon = 0.9$ )	10.69432 $\pm$ 0.1191	2913988.7500 $\pm$ 18332.2363	0.0%	0.0392 $\pm$ 0.0250	9947.0000 $\pm$ 6859.7258	1.8766 $\pm$ 1.3522	512730.0000 $\pm$ 307422.6954	100.0%	64/1052
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	30.90303 $\pm$ 0.2059	8716611.8600 $\pm$ 19277.6545	0.0%	0.1126 $\pm$ 0.1084	31658.2200 $\pm$ 30581.5958	3.5574 $\pm$ 1.1094	999950.0000 $\pm$ 313103.3176	100.0%	64/70
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	30.89782 $\pm$ 0.2076	8711546.9300 $\pm$ 17111.6194	0.0%	0.0979 $\pm$ 0.0971	27449.0200 $\pm$ 27265.8711	2.6415 $\pm$ 0.7523	741970.0000 $\pm$ 211201.6314	100.0%	64/68
Q-learning ( $\rho = 0.9, \epsilon = 1$ )	30.84969 $\pm$ 0.1420	8716392.9200 $\pm$ 10471.9515	0.0%	0.0793 $\pm$ 0.0704	23881.9100 $\pm$ 20016.9771	2.4499 $\pm$ 0.9134	690910.0000 $\pm$ 257280.1623	100.0%	64/510
Q-learning ( $\rho = 0.99, \epsilon = 1$ )	30.94423 $\pm$ 0.1896	8716456.7700 $\pm$ 18100.7792	0.0%	0.1120 $\pm$ 0.1162	31453.8200 $\pm$ 32365.3818	2.4456 $\pm$ 1.0780	687030.0000 $\pm$ 303158.6533	100.0%	64/610
Async Value Iteration	0.24640 $\pm$ 0.0036	212106.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	64/69
Value Iteration	0.40702 $\pm$ 0.0028	344288.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	64/70

**Table 10.** Performance of DP methods and Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.999$  (Problem 10).

Algorithm (Problem 10) ( $\gamma = 0.999$ )	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	1.06206 $\pm$ 0.0155	265867.3800 $\pm$ 75.5967	0.0%	0.0118 $\pm$ 0.0029	2910.3800 $\pm$ 728.7117	0.4371 $\pm$ 0.0062	108280.0000 $\pm$ 601.3319	100.0%	64/48894
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	0.97940 $\pm$ 0.0126	243576.2300 $\pm$ 1468.6347	0.0%	0.0101 $\pm$ 0.0016	2489.1400 $\pm$ 385.4470	0.3118 $\pm$ 0.0046	77440.0000 $\pm$ 682.9348	100.0%	64/8716
Q-learning ( $\rho = 0.9, \epsilon = 0$ )	0.95135 $\pm$ 0.0260	237506.9900 $\pm$ 2795.6508	0.0%	0.0102 $\pm$ 0.0015	2513.4300 $\pm$ 374.4821	0.2440 $\pm$ 0.0167	60840.0000 $\pm$ 4251.3998	100.0%	64/100002
Q-learning ( $\rho = 0.99, \epsilon = 0$ )	0.94648 $\pm$ 0.0174	236133.6700 $\pm$ 2755.9321	0.0%	0.0099 $\pm$ 0.0017	2432.8900 $\pm$ 408.7694	0.2290 $\pm$ 0.0314	57060.0000 $\pm$ 7900.6329	100.0%	64/27731
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	1.38075 $\pm$ 0.0168	348458.2300 $\pm$ 550.3162	0.0%	0.0113 $\pm$ 0.0067	2844.8200 $\pm$ 1667.9400	0.4796 $\pm$ 0.0074	120530.0000 $\pm$ 921.4662	100.0%	64/66
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	1.30497 $\pm$ 0.0156	328004.3100 $\pm$ 553.7289	0.0%	0.0109 $\pm$ 0.0067	2743.1800 $\pm$ 1696.3778	0.3480 $\pm$ 0.0050	87900.0000 $\pm$ 754.9834	100.0%	64/66
Q-learning ( $\rho = 0.9, \epsilon = 0.25$ )	1.25100 $\pm$ 0.0116	317322.7300 $\pm$ 578.4866	0.0%	0.0097 $\pm$ 0.0054	2449.5400 $\pm$ 1376.1401	0.2732 $\pm$ 0.0035	69630.0000 $\pm$ 673.1270	100.0%	64/66
Q-learning ( $\rho = 0.99, \epsilon = 0.25$ )	1.24982 $\pm$ 0.0203	314692.3800 $\pm$ 649.2282	0.0%	0.0104 $\pm$ 0.0063	2622.7000 $\pm$ 1574.8014	0.2513 $\pm$ 0.0056	68510.0000 $\pm$ 754.9172	100.0%	64/66
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	2.01510 $\pm$ 0.0174	525541.2100 $\pm$ 1416.2376	0.0%	0.0165 $\pm$ 0.0087	4313.6500 $\pm$ 2250.0812	0.5370 $\pm$ 0.0081	140430.0000 $\pm$ 1544.3769	100.0%	64/66
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	1.95291 $\pm$ 0.0191	506790.5600 $\pm$ 1395.7975	0.0%	0.0145 $\pm$ 0.0084	3800.7300 $\pm$ 2189.7840	0.4021 $\pm$ 0.0068	105050.0000 $\pm$ 1283.5498	100.0%	64/66
Q-learning ( $\rho = 0.9, \epsilon = 0.5$ )	1.92356 $\pm$ 0.0206	498905.5000 $\pm$ 1280.5515	0.0%	0.0134 $\pm$ 0.0077	3562.3900 $\pm$ 2011.4538	0.3275 $\pm$ 0.0065	85460.0000 $\pm$ 1169.7863	100.0%	64/66
Q-learning ( $\rho = 0.99, \epsilon = 0.5$ )	1.91614 $\pm$ 0.0188	495756.6300 $\pm$ 1433.7520	0.0%	0.0147 $\pm$ 0.0070	3830.3700 $\pm$ 1822.5697	0.3029 $\pm$ 0.0058	79180.0000 $\pm$ 1208.8578	100.0%	64/67
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	1.79071 $\pm$ 0.0309	479534.0000 $\pm$ 89679.4280	100.0%	0.0234 $\pm$ 0.0129	6311.8400 $\pm$ 3458.1080	0.6708 $\pm$ 0.0162	180730.0000 $\pm$ 3709.0565	100.0%	64/66
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	1.35937 $\pm$ 0.0238	364887.2700 $\pm$ 62023.3435	100.0%	0.0243 $\pm$ 0.0137	6523.8400 $\pm$ 3672.1303	0.5161 $\pm$ 0.0099	139210.0000 $\pm$ 2173.9135	100.0%	64/66
Q-learning ( $\rho = 0.9, \epsilon = 0.75$ )	1.07579 $\pm$ 0.0208	286890.2300 $\pm$ 54993.0725	100.0%	0.0243 $\pm$ 0.0130	6590.8100 $\pm$ 3462.0776	0.4383 $\pm$ 0.0111	117390.0000 $\pm$ 2693.4400	100.0%	64/66
Q-learning ( $\rho = 0.99, \epsilon = 0.75$ )	0.97765 $\pm$ 0.0211	262076.3900 $\pm$ 56073.2616	100.0%	0.0251 $\pm$ 0.0131	6727.4000 $\pm$ 3498.1394	0.4122 $\pm$ 0.0107	110750.0000 $\pm$ 2511.4717	100.0%	64/66
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	1.04660 $\pm$ 0.0937	284588.2100 $\pm$ 24490.8895	100.0%	0.0357 $\pm$ 0.0210	9731.1100 $\pm$ 5716.6968	0.8668 $\pm$ 0.0338	235920.0000 $\pm$ 7632.4046	100.0%	64/66
Q-learning ( $\rho = 0.7, \epsilon = 0.9$ )	0.83691 $\pm$ 0.0952	229701.4300 $\pm$ 25612.8380	100.0%	0.0368 $\pm$ 0.0228	10096.6300 $\pm$ 6266.5153	0.6819 $\pm$ 0.0236	187190.0000 $\pm$ 6042.6732	100.0%	64/66
Q-learning ( $\rho = 0.9, \epsilon = 0.9$ )	0.71331 $\pm$ 0.0724	194477.8400 $\pm$ 19500.5225	100.0%	0.0375 $\pm$ 0.0232	10217.5800 $\pm$ 6392.4782	0.5952 $\pm$ 0.0212	162210.0000 $\pm$ 5332.2599	100.0%	64/67
Q-learning ( $\rho = 0.99, \epsilon = 0.9$ )	0.65742 $\pm$ 0.0572	180192.3600 $\pm$ 15260.4553	100.0%	0.0371 $\pm$ 0.0219	10147.8100 $\pm$ 6017.0427	0.5611 $\pm$ 0.0225	153560.0000 $\pm$ 5817.7659	100.0%	64/66
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	4.84662 $\pm$ 1.4434	1342809.3100 $\pm$ 399317.5826	100.0%	0.1142 $\pm$ 0.1162	31313.3500 $\pm$ 31759.6827	2.9789 $\pm$ 0.5718	824560.0000 $\pm$ 157018.8091	100.0%	64/66
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	3.52291 $\pm$ 0.9998	985856.3500 $\pm$ 282047.3120	100.0%	0.0975 $\pm$ 0.0970	27210.7400 $\pm$ 27062.4015	2.2450 $\pm$ 0.4430	627880.0000 $\pm$ 124366.2559	100.0%	64/68
Q-learning ( $\rho = 0.9, \epsilon = 1$ )	2.80282 $\pm$ 1.0162	777332.6400 $\pm$ 281713.9771	100.0%	0.0900 $\pm$ 0.0809	24856.4400 $\pm$ 22392.3595	1.8761 $\pm$ 0.4569	519830.0000 $\pm$ 125921.3290	100.0%	64/67
Q-learning ( $\rho = 0.99, \epsilon = 1$ )	2.82997 $\pm$ 0.7142	629074.5400 $\pm$ 190975.3965	100.0%	0.0972 $\pm$ 0.0793	26692.6900 $\pm$ 21712.2338	1.6362 $\pm$ 0.4014	449470.0000 $\pm$ 110564.1102	100.0%	64/66
Async Value Iteration	0.21926 $\pm$ 0.0162	178292.0000 $\pm$ 0.0000	N/A	N/A	N/A	N			

**Table 11.** Performance of DP methods and Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.99$  (Problem 11).

Algorithm (Problem 11) ( $\gamma = 0.99$ )	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	0.39799 $\pm$ 0.0210	95423.6800 $\pm$ 567.6079	0.0%	0.0067 $\pm$ 0.0011	1633.4300 $\pm$ 258.5025	0.1518 $\pm$ 0.0101	36850.0000 $\pm$ 2321.0989	100.0%	22/10002
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	0.37312 $\pm$ 0.0115	96568.8000 $\pm$ 2163.3390	0.0%	0.0067 $\pm$ 0.0011	1637.7000 $\pm$ 266.1527	0.1241 $\pm$ 0.0323	30404.0000 $\pm$ 7841.7675	99.0%	22/310
Q-learning ( $\rho = 0.9, \epsilon = 0$ )	0.38073 $\pm$ 0.0124	92259.2500 $\pm$ 2742.9887	0.0%	0.0068 $\pm$ 0.0010	1675.1000 $\pm$ 256.6727	0.1323 $\pm$ 0.0711	32383.5616 $\pm$ 17337.6067	73.0%	22/3046
Q-learning ( $\rho = 0.99, \epsilon = 0$ )	0.38594 $\pm$ 0.0130	93504.7500 $\pm$ 2852.7542	0.0%	0.0065 $\pm$ 0.0008	1590.2000 $\pm$ 183.6596	0.1885 $\pm$ 0.0765	45947.3684 $\pm$ 18453.2801	38.0%	22/606
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	0.51791 $\pm$ 0.0067	126854.4000 $\pm$ 347.8654	0.0%	0.0080 $\pm$ 0.0018	1994.0800 $\pm$ 454.6913	0.1593 $\pm$ 0.0079	39700.0000 $\pm$ 1920.9373	100.0%	22/27
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	0.49021 $\pm$ 0.0080	119017.8200 $\pm$ 575.3976	0.0%	0.0074 $\pm$ 0.0015	1833.9600 $\pm$ 376.2915	0.1308 $\pm$ 0.0220	32300.0000 $\pm$ 5243.0907	100.0%	22/26
Q-learning ( $\rho = 0.9, \epsilon = 0.25$ )	0.47611 $\pm$ 0.0072	115911.2900 $\pm$ 736.9714	0.0%	0.0075 $\pm$ 0.0017	1864.9000 $\pm$ 422.5565	0.1425 $\pm$ 0.0551	35121.2121 $\pm$ 13224.0088	99.0%	22/82
Q-learning ( $\rho = 0.99, \epsilon = 0.25$ )	0.47157 $\pm$ 0.0060	115410.4400 $\pm$ 700.0474	0.0%	0.0074 $\pm$ 0.0017	1838.1700 $\pm$ 438.8607	0.2307 $\pm$ 0.1067	56633.3333 $\pm$ 25843.7398	30.0%	22/2063
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	0.76524 $\pm$ 0.0090	192499.1600 $\pm$ 812.2625	0.0%	0.0083 $\pm$ 0.0030	2131.3800 $\pm$ 781.6668	0.1784 $\pm$ 0.0105	45770.0000 $\pm$ 2591.7369	100.0%	22/25
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	0.73667 $\pm$ 0.0096	185125.1100 $\pm$ 827.9211	0.0%	0.0083 $\pm$ 0.0028	2118.1700 $\pm$ 725.6840	0.1429 $\pm$ 0.0185	36590.0000 $\pm$ 4630.5399	100.0%	22/25
Q-learning ( $\rho = 0.9, \epsilon = 0.5$ )	0.72997 $\pm$ 0.0089	181844.2000 $\pm$ 870.7109	0.0%	0.0083 $\pm$ 0.0028	2129.4300 $\pm$ 740.6331	0.1584 $\pm$ 0.0545	40860.0000 $\pm$ 13635.6298	100.0%	22/1680
Q-learning ( $\rho = 0.99, \epsilon = 0.5$ )	0.72236 $\pm$ 0.0078	181875.7900 $\pm$ 944.6629	0.0%	0.0088 $\pm$ 0.0025	2245.8600 $\pm$ 653.2801	0.3493 $\pm$ 0.1707	88333.3333 $\pm$ 42697.5335	21.0%	22/2313
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	1.55484 $\pm$ 0.0213	404114.4100 $\pm$ 2614.4949	0.0%	0.0112 $\pm$ 0.0039	2974.8200 $\pm$ 1040.4007	0.2188 $\pm$ 0.0166	57790.0000 $\pm$ 4393.4497	100.0%	22/25
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	1.53926 $\pm$ 0.0216	400345.6600 $\pm$ 3194.0362	0.0%	0.0099 $\pm$ 0.0047	2622.3500 $\pm$ 1254.2456	0.1835 $\pm$ 0.0228	48470.0000 $\pm$ 6104.8423	100.0%	22/25
Q-learning ( $\rho = 0.9, \epsilon = 0.75$ )	1.54499 $\pm$ 0.0296	402485.0100 $\pm$ 3280.6152	0.0%	0.0107 $\pm$ 0.0040	2919.8400 $\pm$ 1074.0321	0.1793 $\pm$ 0.0451	47290.0000 $\pm$ 11888.8982	100.0%	22/25
Q-learning ( $\rho = 0.99, \epsilon = 0.75$ )	1.57677 $\pm$ 0.0174	411567.1500 $\pm$ 2845.5532	0.0%	0.0101 $\pm$ 0.0040	2689.6600 $\pm$ 1072.7643	0.2794 $\pm$ 0.4547	190827.5862 $\pm$ 118468.1156	29.0%	22/25
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	4.06947 $\pm$ 0.0061	1089813.3000 $\pm$ 11695.3669	0.0%	0.0101 $\pm$ 0.0056	2707.5800 $\pm$ 1508.7474	0.2907 $\pm$ 0.0359	78120.0000 $\pm$ 9666.7264	100.0%	22/24
Q-learning ( $\rho = 0.7, \epsilon = 0.9$ )	4.05351 $\pm$ 0.0059	1091703.0000 $\pm$ 12073.2073	0.0%	0.0095 $\pm$ 0.0057	2799.7000 $\pm$ 1775.6700	0.0793 $\pm$ 0.0469	65000.0000 $\pm$ 10853.5798	100.0%	22/25
Q-learning ( $\rho = 0.9, \epsilon = 0.9$ )	4.29611 $\pm$ 0.0819	1124205.1000 $\pm$ 12992.2186	0.0%	0.0097 $\pm$ 0.0055	2599.7100 $\pm$ 1492.5060	0.2830 $\pm$ 0.0883	75330.0000 $\pm$ 23466.1693	100.0%	22/3000
Q-learning ( $\rho = 0.99, \epsilon = 0.9$ )	4.51425 $\pm$ 0.0887	1198876.9800 $\pm$ 14514.8737	0.0%	0.0089 $\pm$ 0.0053	2387.6900 $\pm$ 1443.7231	0.2070 $\pm$ 0.1219	548638.2979 $\pm$ 320207.8066	47.0%	22/300
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	17.38310 $\pm$ 0.2022	4830298.6800 $\pm$ 48616.3395	0.0%	0.0072 $\pm$ 0.0067	2157.7200 $\pm$ 1850.2128	0.4759 $\pm$ 0.0839	132370.0000 $\pm$ 23297.4913	100.0%	22/26
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	17.52310 $\pm$ 0.2141	4828398.3300 $\pm$ 52080.5831	0.0%	0.0073 $\pm$ 0.0064	2099.8400 $\pm$ 1783.1210	0.4094 $\pm$ 0.1053	113350.0000 $\pm$ 29966.6045	100.0%	22/25
Q-learning ( $\rho = 0.9, \epsilon = 1$ )	17.54003 $\pm$ 0.2390	4841830.8000 $\pm$ 54715.3973	0.0%	0.0066 $\pm$ 0.0056	1799.5900 $\pm$ 1538.6349	0.4463 $\pm$ 0.1492	123170.0000 $\pm$ 41225.4909	100.0%	22/26
Q-learning ( $\rho = 0.99, \epsilon = 1$ )	17.47704 $\pm$ 0.2397	4828782.2700 $\pm$ 54208.8381	0.0%	0.0064 $\pm$ 0.0053	1769.1100 $\pm$ 1467.0934	0.3585 $\pm$ 0.4520	2033564.5161 $\pm$ 1253221.1480	62.0%	22/26
Async Value Iteration	0.20147 $\pm$ 0.0022	178048.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	22/25
Value Iteration	0.27400 $\pm$ 0.0051	232832.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	22/25

**Table 12.** Performance of DP methods and Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.999$  (Problem 11).

Algorithm (Problem 11) ( $\gamma = 0.999$ )	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	0.38377 $\pm$ 0.0045	93082.8800 $\pm$ 193.1769	0.0%	0.0056 $\pm$ 0.0004	1375.0100 $\pm$ 86.5073	0.1436 $\pm$ 0.0018	35020.0000 $\pm$ 140.0000	100.0%	22/26
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	0.35480 $\pm$ 0.0113	85400.1200 $\pm$ 660.9158	0.0%	0.0056 $\pm$ 0.0005	1372.6300 $\pm$ 108.3105	0.1035 $\pm$ 0.0017	25130.0000 $\pm$ 364.8287	100.0%	22/4025
Q-learning ( $\rho = 0.9, \epsilon = 0$ )	0.34218 $\pm$ 0.0090	82557.8800 $\pm$ 1606.2870	0.0%	0.0056 $\pm$ 0.0003	1357.6200 $\pm$ 65.3582	0.0828 $\pm$ 0.0021	20160.0000 $\pm$ 417.6123	100.0%	22/21324
Q-learning ( $\rho = 0.99, \epsilon = 0$ )	0.34383 $\pm$ 0.0101	82572.4900 $\pm$ 1944.7986	0.0%	0.0056 $\pm$ 0.0003	1360.0700 $\pm$ 77.6700	0.0793 $\pm$ 0.0029	21370.0000 $\pm$ 630.9231	100.0%	22/8643
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	0.51539 $\pm$ 0.0073	124786.3700 $\pm$ 373.1031	0.0%	0.0080 $\pm$ 0.0019	1971.3500 $\pm$ 464.3022	0.1532 $\pm$ 0.0029	37780.0000 $\pm$ 414.2463	100.0%	22/24
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	0.47911 $\pm$ 0.0046	116563.4500 $\pm$ 358.9886	0.0%	0.0075 $\pm$ 0.0016	1861.1700 $\pm$ 391.1360	0.1129 $\pm$ 0.0013	28010.0000 $\pm$ 223.3831	100.0%	22/24
Q-learning ( $\rho = 0.9, \epsilon = 0.25$ )	0.46427 $\pm$ 0.0052	112396.8000 $\pm$ 394.4297	0.0%	0.0077 $\pm$ 0.0013	1895.2900 $\pm$ 325.2237	0.0925 $\pm$ 0.0021	22780.0000 $\pm$ 181.2484	100.0%	22/23
Q-learning ( $\rho = 0.99, \epsilon = 0.25$ )	0.46423 $\pm$ 0.0088	111608.8200 $\pm$ 385.9230	0.0%	0.0075 $\pm$ 0.0015	1836.5800 $\pm$ 373.8410	0.0872 $\pm$ 0.0028	21370.0000 $\pm$ 204.2222	100.0%	22/24
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	0.76484 $\pm$ 0.0100	189923.3500 $\pm$ 1146.3437	0.0%	0.0086 $\pm$ 0.0024	2169.1000 $\pm$ 609.7160	0.1696 $\pm$ 0.0030	43010.0000 $\pm$ 538.4296	100.0%	22/22
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	0.73026 $\pm$ 0.0113	183074.4400 $\pm$ 1037.7751	0.0%	0.0081 $\pm$ 0.0027	2074.3600 $\pm$ 687.7962	0.1280 $\pm$ 0.0031	32700.0000 $\pm$ 591.6080	100.0%	22/24
Q-learning ( $\rho = 0.9, \epsilon = 0.5$ )	0.71527 $\pm$ 0.0085	180180.1500 $\pm$ 1102.2192	0.0%	0.0080 $\pm$ 0.0023	2046.0700 $\pm$ 605.3878	0.1064 $\pm$ 0.0027	27270.0000 $\pm$ 597.5784	100.0%	22/22
Q-learning ( $\rho = 0.99, \epsilon = 0.5$ )	0.71556 $\pm$ 0.0084	181054.6100 $\pm$ 955.8296	0.0%	0.0082 $\pm$ 0.0022	2130.3800 $\pm$ 580.1694	0.0993 $\pm$ 0.0029	25570.0000 $\pm$ 630.9192	100.0%	22/23
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	1.58401 $\pm$ 0.0221	415905.4000 $\pm$ 5266.6597	0.0%	0.0103 $\pm$ 0.0038	2729.0300 $\pm$ 1019.6230	0.2052 $\pm$ 0.0054	54160.0000 $\pm$ 1293.9861	100.0%	22/23
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	1.57483 $\pm$ 0.0245	409969.3600 $\pm$ 4411.5591	0.0%	0.0095 $\pm$ 0.0041	2505.6300 $\pm$ 1089.8314	0.1598 $\pm$ 0.0045	42230.0000 $\pm$ 1156.3304	100.0%	22/24
Q-learning ( $\rho = 0.9, \epsilon = 0.75$ )	1.59856 $\pm$ 0.0238	417749.4400 $\pm$ 5188.3083	0.0%	0.0109 $\pm$ 0.0040	2883.3000 $\pm$ 1064.7107	0.1368 $\pm$ 0.0054	36210.0000 $\pm$ 1351.2587	100.0%	22/24
Q-learning ( $\rho = 0.99, \epsilon = 0.75$ )	1.62679 $\pm$ 0.2026	433952.5000 $\pm$ 1194.3870	0.0%	0.0096 $\pm$ 0.0038	2834.7000 $\pm$ 1077.4538	0.1280 $\pm$ 0.0040	34110.0000 $\pm$ 908.9494	100.0%	22/23
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	1.00714 $\pm$ 0.1571	269682.8100 $\pm$ 41926.1143	100.0%	0.0100 $\pm$ 0.0060	2708.0700 $\pm$ 1615.6574	0.2630 $\pm$ 0.0117	70880.0000 $\pm$ 2997.5990	100.0%	22/24
Q-learning ( $\rho = 0.7, \epsilon = 0.9$ )	0.75531 $\pm$ 0.1075	200773.1600 $\pm$ 28429.8556	100.0%	0.0102 $\pm$ 0.0058	2714.7700 $\pm$ 1565.0885	0.2099 $\pm$ 0.0123	56100.0000 $\pm$ 3106.4449	100.0%	22/24
Q-learning ( $\rho = 0.9, \epsilon = 0.9$ )	0.70365 $\pm$ 0.1405	184876.9200 $\pm$ 36768.9518	100.0%	0.0098 $\pm$ 0.0056	2589.9500 $\pm$ 1502.0573	0.1851 $\pm$ 0.0198	49920.0000 $\pm$ 2777.3369	100.0%	22/24
Q-learning ( $\rho = 0.99, \epsilon = 0.9$ )	0.65325 $\pm$ 0.1262	160641.6100 $\pm$ 32985.1712	100.0%	0.0108 $\pm$ 0.0061	2890.6600 $\pm$ 1640.6114	0.1714 $\pm$ 0.0078	45720.0000 $\pm$ 2040.0000	100.0%	22/23
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	0.87511 $\pm$ 0.1260	241990.0900 $\pm$ 34516.0792	100.0%	0.0075 $\pm$ 0.0064	2077.2500 $\pm$ 1780.5063	0.4522 $\pm$ 0.0631	125150.0000 $\pm$ 17723.0782	100.0%	22/23
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	0.70148 $\pm$ 0.0778	193154.4900 $\pm$ 21280.4327	100.0%	0.0073 $\pm$ 0.0061	2000.0100 $\pm$ 1683.8940	0.3587 $\pm$ 0.0626	98660.0000 $\pm$ 17117.9555	100.0%	22/24
Q-learning ( $\rho = 0.9, \epsilon = 1$ )	0.60545 $\pm$ 0.0850	167396.2800 $\pm$ 23265.3554	100.0%	0.0072 $\pm$ 0.0059	1964.9300 $\pm$ 1621.4139	0.3087 $\pm$ 0.0572	85260.0000 $\pm$ 15863.4013	100.0%	22/22
Q-learning ( $\rho = 0.99, \epsilon = 1$ )	0.63211 $\pm$ 0.0721	1483071.8100 $\pm$ 20421.1873	100.0%	0.0066 $\pm$ 0.0051	1779.6300 $\pm$ 1419.8986	0.2624 $\pm$ 0.0442	73390.0000 $\pm$ 12284.0506	100.0%	22/23
Async Value Iteration	0.16458 $\pm$ 0.0020	147232.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	22/23
Value Iteration	0.21802 $\pm$ 0.0008	184896.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	22/23

**Table 14.** Performance of DP methods and Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.7$  (Problem 1).

Algorithm (Problem 1) ( $\gamma = 0.7$ )	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.05, \epsilon = 0$ )	1.93393 $\pm$ 0.0889	518018.8000 $\pm$ 1079.4022	0.0%	0.0082 $\pm$ 0.0027	1755.0000 $\pm$ 772.9291	nan $\pm$ nan	nan $\pm$ nan	0.0%	37/312
Q-learning ( $\rho = 0.1, \epsilon = 0$ )	1.22601 $\pm$ 0.0191	346009.9000 $\pm$ 1428.1871	0.0%	0.0086 $\pm$ 0.0030	2369.8000 $\pm$ 840.9631	1.2029 $\pm$ 0.0272	340500.0000 $\pm$ 4193.2485	60.0%	35/92
Q-learning ( $\rho = 0.3, \epsilon = 0$ )	0.75712 $\pm$ 0.0197	211921.6000 $\pm$ 1266.0164	0.0%	0.0059 $\pm$ 0.0026	1675.6000 $\pm$ 592.5429	0.4032 $\pm$ 0.0182	113700.0000 $\pm$ 3848.3763	100.0%	36/49
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	0.65779 $\pm$ 0.0115	186711.0000 $\pm$ 1534.7377	0.0%	0.0084 $\pm$ 0.0036	2143.8000 $\pm$ 1069.8904	0.2517 $\pm$ 0.0298	70900.0000 $\pm$ 8129.5756	100.0%	41/56
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	0.63883 $\pm$ 0.0130	183386.4000 $\pm$ 4281.1417	0.0%	0.0061 $\pm$ 0.0021	1684.2000 $\pm$ 597.2734	0.2239 $\pm$ 0.0784	64200.0000 $\pm$ 21469.9790	100.0%	35/70
Q-learning ( $\rho = 0.05, \epsilon = 0.25$ )	1.84066 $\pm$ 0.0490	532172.3000 $\pm$ 1821.7158	0.0%	0.0074 $\pm$ 0.0034	2220.0000 $\pm$ 1046.5392	nan $\pm$ nan	nan $\pm$ nan	0.0%	41/312
Q-learning ( $\rho = 0.1, \epsilon = 0.25$ )	1.28438 $\pm$ 0.0172	371055.9000 $\pm$ 800.9085	0.0%	0.0082 $\pm$ 0.0060	2265.8000 $\pm$ 1483.7717	1.2825 $\pm$ 0.0000	371000.0000 $\pm$ 0.0000	10.0%	34/55
Q-learning ( $\rho = 0.3, \epsilon = 0.25$ )	0.88966 $\pm$ 0.0169	258582.0000 $\pm$ 1501.4790	0.0%	0.0075 $\pm$ 0.0019	2033.1000 $\pm$ 549.4465	0.5150 $\pm$ 0.0621	151000.0000 $\pm$ 19271.7410	100.0%	32/98
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	0.80985 $\pm$ 0.0169	254281.8000 $\pm$ 690.4668	0.0%	0.0082 $\pm$ 0.0030	2355.4000 $\pm$ 835.7866	0.3277 $\pm$ 0.0637	96200.0000 $\pm$ 18258.1489	100.0%	39/73
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	0.96430 $\pm$ 0.0235	279320.7000 $\pm$ 2157.6759	0.0%	0.0076 $\pm$ 0.0025	2163.3000 $\pm$ 735.2480	0.2296 $\pm$ 0.0497	66000.0000 $\pm$ 14696.9385	100.0%	39/111
Q-learning ( $\rho = 0.05, \epsilon = 0.5$ )	1.95859 $\pm$ 0.0273	586129.8000 $\pm$ 2392.9607	0.0%	0.0074 $\pm$ 0.0028	2343.3000 $\pm$ 827.6855	nan $\pm$ nan	nan $\pm$ nan	0.0%	35/70
Q-learning ( $\rho = 0.1, \epsilon = 0.5$ )	1.50248 $\pm$ 0.0178	473398.7000 $\pm$ 2670.3218	0.0%	0.0086 $\pm$ 0.0037	2991.9000 $\pm$ 1129.5691	1.3737 $\pm$ 0.0000	412900.0000 $\pm$ 0.0000	10.0%	33/48
Q-learning ( $\rho = 0.3, \epsilon = 0.5$ )	1.20397 $\pm$ 0.0226	358492.8000 $\pm$ 1388.4665	0.0%	0.0086 $\pm$ 0.0056	2532.9000 $\pm$ 1536.8264	0.5031 $\pm$ 0.0000	155000.0000 $\pm$ 0.0000	10.0%	36/51
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	1.23370 $\pm$ 0.0241	366895.6000 $\pm$ 2016.7670	0.0%	0.0084 $\pm$ 0.0029	2662.9000 $\pm$ 776.2440	0.7441 $\pm$ 0.2714	223333.3333 $\pm$ 80400.3870	30.0%	39/135
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	1.36602 $\pm$ 0.0202	405297.1000 $\pm$ 2158.8903	0.0%	0.0095 $\pm$ 0.0041	2778.8000 $\pm$ 1035.1374	0.7672 $\pm$ 0.4735	226000.0000 $\pm$ 135000.0000	20.0%	49/206
Q-learning ( $\rho = 0.05, \epsilon = 0.75$ )	2.63461 $\pm$ 0.0437	804658.0000 $\pm$ 4958.7344	0.0%	0.0077 $\pm$ 0.0039	2453.1000 $\pm$ 1264.9290	nan $\pm$ nan	nan $\pm$ nan	0.0%	34/66
Q-learning ( $\rho = 0.1, \epsilon = 0.75$ )	2.23579 $\pm$ 0.0544	714200.6000 $\pm$ 4243.7842	0.0%	0.0083 $\pm$ 0.0040	2721.3000 $\pm$ 1333.0644	nan $\pm$ nan	nan $\pm$ nan	0.0%	35/54
Q-learning ( $\rho = 0.3, \epsilon = 0.75$ )	2.09277 $\pm$ 0.0412	671241.6000 $\pm$ 4798.8179	0.0%	0.0104 $\pm$ 0.0041	3375.1000 $\pm$ 1409.1686	nan $\pm$ nan	nan $\pm$ nan	0.0%	37/91
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	2.20245 $\pm$ 0.0270	702471.4000 $\pm$ 6924.6805	0.0%	0.0101 $\pm$ 0.0052	3063.8000 $\pm$ 1475.5833	nan $\pm$ nan	nan $\pm$ nan	0.0%	41/90
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	2.40040 $\pm$ 0.0310	778859.2000 $\pm$ 5927.0231	0.0%	0.0087 $\pm$ 0.0052	2838.4000 $\pm$ 1084.1625	nan $\pm$ nan	nan $\pm$ nan	0.0%	49/216
Q-learning ( $\rho = 0.05, \epsilon = 0.9$ )	4.63566 $\pm$ 0.0648	1529885.7000 $\pm$ 14965.9439	0.0%	0.0093 $\pm$ 0.0054	2950.3000 $\pm$ 1811.1419	nan $\pm$ nan	nan $\pm$ nan	0.0%	38/59
Q-learning ( $\rho = 0.1, \epsilon = 0.9$ )	4.53196 $\pm$ 0.0503	1504314.0000 $\pm$ 17157.8236	0.0%	0.0084 $\pm$ 0.0054	2891.9000 $\pm$ 1814.5724	nan $\pm$ nan	nan $\pm$ nan	0.0%	32/49
Q-learning ( $\rho = 0.3, \epsilon = 0.9$ )	4.69049 $\pm$ 0.1418	1518928.1000 $\pm$ 16357.8860	0.0%	0.0100 $\pm$ 0.0074	3207.4000 $\pm$ 2346.9705	nan $\pm$ nan	nan $\pm$ nan	0.0%	36/54
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	5.18138 $\pm$ 0.2839	1606067.0000 $\pm$ 14323.8288	0.0%	0.0106 $\pm$ 0.0074	2953.4000 $\pm$ 1984.1909	nan $\pm$ nan	nan $\pm$ nan	0.0%	39/67
Q-learning ( $\rho = 0.7, \epsilon = 0.9$ )	5.32480 $\pm$ 0.0771	1747653.4000 $\pm$ 18328.3763	0.0%	0.0081 $\pm$ 0.0044	2636.9000 $\pm$ 1530.3755	nan $\pm$ nan	nan $\pm$ nan	0.0%	43/213
Q-learning ( $\rho = 0.05, \epsilon = 1$ )	14.37472 $\pm$ 0.1425	4828867.8000 $\pm$ 36170.1378	0.0%	0.0061 $\pm$ 0.0060	1831.1000 $\pm$ 2147.7063	nan $\pm$ nan	nan $\pm$ nan	0.0%	34/51
Q-learning ( $\rho = 0.1, \epsilon = 1$ )	14.45536 $\pm$ 0.2361	4827837.3000 $\pm$ 54592.7340	0.0%	0.0062 $\pm$ 0.0037	1993.1000 $\pm$ 1234.6338	nan $\pm$ nan	nan $\pm$ nan	0.0%	34/48
Q-learning ( $\rho = 0.3, \epsilon = 1$ )	14.30180 $\pm$ 0.1002	4806959.1000 $\pm$ 32144.8814	0.0%	0.0041 $\pm$ 0.0041	1327.7000 $\pm$ 1298.6299	nan $\pm$ nan	nan $\pm$ nan	0.0%	36/53
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	14.39015 $\pm$ 0.1338	4811996.4000 $\pm$ 32298.9266	0.0%	0.0059 $\pm$ 0.0069	1954.0000 $\pm$ 2181.1991	nan $\pm$ nan	nan $\pm$ nan	0.0%	34/76
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	14.34809 $\pm$ 0.1401	4807284.1000 $\pm$ 33911.3460	0.0%	0.0053 $\pm$ 0.0033	1504.8000 $\pm$ 1073.6604	nan $\pm$ nan	nan $\pm$ nan	0.0%	35/90
Asyn Value Iteration	0.33047 $\pm$ 0.0064	473424.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	38/50
Value Iteration	0.50604 $\pm$ 0.0121	721728.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	32/62

**Table 15.** Performance of DP methods and Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.9$  (Problem 1).

Algorithm (Problem 1) ( $\gamma = 0.9$ )	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.05, \epsilon = 0$ )	1.40630 $\pm$ 0.0248	374388.4000 $\pm$ 465.3979	0.0%	0.0076 $\pm$ 0.0010	1905.9000 $\pm$ 403.3647	nan $\pm$ nan	nan $\pm$ nan	0.0%	36/1538
Q-learning ( $\rho = 0.1, \epsilon = 0$ )	0.94216 $\pm$ 0.0373	245359.1000 $\pm$ 529.2448	0.0%	0.0058 $\pm$ 0.0020	1530.3000 $\pm$ 411.9168	0.9288 $\pm$ 0.0408	241000.0000 $\pm$ 2236.0680	60.0%	29/149
Q-learning ( $\rho = 0.3, \epsilon = 0$ )	0.57358 $\pm$ 0.0309	146838.2000 $\pm$ 505.2702	0.0%	0.0052 $\pm$ 0.0013	1359.0000 $\pm$ 472.7898	0.3277 $\pm$ 0.0246	83100.0000 $\pm$ 3419.0642	100.0%	28/46
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	0.48362 $\pm$ 0.0215	128080.1000 $\pm$ 676.8250	0.0%	0.0054 $\pm$ 0.0015	1342.8000 $\pm$ 305.3411	0.1893 $\pm$ 0.0185	48900.0000 $\pm$ 1700.0000	100.0%	28/40
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	0.46465 $\pm$ 0.0163	122683.3000 $\pm$ 2014.7824	0.0%	0.0043 $\pm$ 0.0028	1179.8000 $\pm$ 420.3148	0.1604 $\pm$ 0.0362	41800.0000 $\pm$ 2166.5891	100.0%	28/37
Q-learning ( $\rho = 0.05, \epsilon = 0.25$ )	1.43774 $\pm$ 0.0240	387669.7000 $\pm$ 747.8249	0.0%	0.0066 $\pm$ 0.0028	1886.5000 $\pm$ 696.6657	nan $\pm$ nan	nan $\pm$ nan	0.0%	30/292
Q-learning ( $\rho = 0.1, \epsilon = 0.25$ )	0.97812 $\pm$ 0.0325	269944.7000 $\pm$ 588.4646	0.0%	0.0082 $\pm$ 0.0017	2255.3000 $\pm$ 486.6432	0.9239 $\pm$ 0.0000	263000.0000 $\pm$ 0.0000	10.0%	28/34
Q-learning ( $\rho = 0.3, \epsilon = 0.25$ )	0.65382 $\pm$ 0.0262	178714.2000 $\pm$ 602.5748	0.0%	0.0062 $\pm$ 0.0009	1630.3000 $\pm$ 348.9653	0.5085 $\pm$ 0.0471	136000.0000 $\pm$ 10977.2492	40.0%	30/39
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	0.60797 $\pm$ 0.0249	162876.3000 $\pm$ 617.9632	0.0%	0.0058 $\pm$ 0.0016	1642.6000 $\pm$ 419.5217	0.3219 $\pm$ 0.0223	85777.7778 $\pm$ 24530.1528	90.0%	30/43
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	0.59050 $\pm$ 0.0233	160202.3000 $\pm$ 757.8191	0.0%	0.0067 $\pm$ 0.0022	1651.8000 $\pm$ 461.5987	0.2097 $\pm$ 0.0428	58200.0000 $\pm$ 10097.5244	100.0%	28/87
Q-learning ( $\rho = 0.05, \epsilon = 0.5$ )	1.53399 $\pm$ 0.0232	429555.3000 $\pm$ 1100.5893	0.0%	0.0073 $\pm$ 0.0026	2106.6000 $\pm$ 748.5252	nan $\pm$ nan	nan $\pm$ nan	0.0%	30/284
Q-learning ( $\rho = 0.1, \epsilon = 0.5$ )	1.14633 $\pm$ 0.0310	324285.5000 $\pm$ 1215.7992	0.0%	0.0080 $\pm$ 0.0027	2126.5000 $\pm$ 517.5400	1.0349 $\pm$ 0.0000	295000.0000 $\pm$ 0.0000	10.0%	29/38
Q-learning ( $\rho = 0.3, \epsilon = 0.5$ )	0.90733 $\pm$ 0.0415	251963.2000 $\pm$ 1651.1753	0.0%	0.0074 $\pm$ 0.0029	2008.2000 $\pm$ 685.7393	nan $\pm$ nan	nan $\pm$ nan	0.0%	28/35
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	0.83558 $\pm$ 0.0293	237662.9000 $\pm$ 1018.6100	0.0%	0.0068 $\pm$ 0.0016	1893.6000 $\pm$ 501.3544	0.8210 $\pm$ 0.0000	236000.0000 $\pm$ 0.0000	10.0%	28/34
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	0.84558 $\pm$ 0.0386	238293.1000 $\pm$ 1387.4264	0.0%	0.0070 $\pm$ 0.0017	1948.5000 $\pm$ 425.8716	0.3798 $\pm$ 0.1223	107250.0000 $\pm$ 33461.7319	40.0%	29/94
Q-learning ( $\rho = 0.05, \epsilon = 0.75$ )	2.06221 $\pm$ 0.0449	596335.3000 $\pm$ 2529.3901	0.0%	0.0098 $\pm$ 0.0042	2738.5000 $\pm$ 1076.0376	nan $\pm$ nan	nan $\pm$ nan	0.0%	30/40
Q-learning ( $\rho = 0.1, \epsilon = 0.75$ )	1.83990 $\pm$ 0.0367	526975.8000 $\pm$ 2997.9857	0.0%	0.0084 $\pm$ 0.0026	2374.0000 $\pm$ 784.2525	nan $\pm$ nan	nan $\pm$ nan	0.0%	29/37
Q-learning ( $\rho = 0.3, \epsilon = 0.75$ )	1.66193 $\pm$ 0.0250	479980.7000 $\pm$ 3110.5655	0.0%	0.0087 $\pm$ 0.0031	2513.5000 $\pm$ 931.2594	nan $\pm$ nan	nan $\pm$ nan	0.0%	28/34
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	1.63547 $\pm$ 0.0554	473658.0000 $\pm$ 2320.7265	0.0%	0.0100 $\pm$ 0.0043	2728.8000 $\pm$ 1042.9370	0.5347 $\pm$ 0.0000	155000.0000 $\pm$ 0.0000	10.0%	29/38
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	1.67548 $\pm$ 0.0434	478412.8000 $\pm$ 2841.9262	0.0%	0.0086 $\pm$ 0.0022	2461.2000 $\pm$ 788.4459	nan $\pm$ nan	nan $\pm$ nan	0.0%	28/87
Q-learning ( $\rho = 0.05, \epsilon = 0.9$ )	3.89532 $\pm$ 0.0531	115420.9000 $\pm$ 10356.6597	0.0%	0.0104 $\pm$ 0.0072	3129.3000 $\pm$ 2040.2814	nan $\pm$ nan	nan $\pm$ nan	0.0%	29/35
Q-learning ( $\rho = 0.1, \epsilon = 0.9$ )	3.83259 $\pm$ 0.0690	113688.5000 $\pm$ 5079.6989	0.0%	0.0098 $\pm$ 0.0053	2640.0000 $\pm$ 1763.6383	nan $\pm$ nan	nan $\pm$ nan	0.0%	29/36
Q-learning ( $\rho = 0.3, \epsilon = 0.9$ )	3.87358 $\pm$ 0.0763	1133814.1000 $\pm$ 8302.0288	0.0%	0.0074 $\pm$ 0.0042	2248.7000 $\pm$ 1257.7827	nan $\pm$ nan	nan $\pm$ nan	0.0%	29/36
Q-learning ( $\rho =$									

**Table 16.** Performance of DP methods and Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.5$  (Problem 10).

Algorithm (Problem 10) ( $\gamma = 0.5$ )	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.05, \epsilon = 0$ )	9.4043 $\pm$ 0.1140	2480472.5000 $\pm$ 7753.3471	0.0%	0.1200 $\pm$ 0.0847	32601.0000 $\pm$ 25738.4843	nan $\pm$ nan	nan $\pm$ nan	0.0%	2387/415
Q-learning ( $\rho = 0.1, \epsilon = 0$ )	6.02790 $\pm$ 0.1197	1620228.2000 $\pm$ 7376.2157	0.0%	0.0539 $\pm$ 0.0348	25728.8000 $\pm$ 9519.7926	5.7811 $\pm$ 0.1068	1553000.0000 $\pm$ 18606.4505	100.0%	163/232
Q-learning ( $\rho = 0.3, \epsilon = 0$ )	3.54922 $\pm$ 0.0715	975451.9000 $\pm$ 9655.4821	0.0%	0.0452 $\pm$ 0.0357	12364.1000 $\pm$ 9482.9381	1.9270 $\pm$ 0.1422	532714.2857 $\pm$ 43839.6894	70.0%	142/263
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	3.17621 $\pm$ 0.0505	879102.4000 $\pm$ 13869.9244	0.0%	0.0405 $\pm$ 0.0282	11513.0000 $\pm$ 7962.2145	1.6239 $\pm$ 0.0521	447666.6667 $\pm$ 17849.4949	30.0%	197/253
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	3.12813 $\pm$ 0.1029	865605.3000 $\pm$ 28493.4129	0.0%	0.0335 $\pm$ 0.0212	9296.6000 $\pm$ 6125.0827	1.0988 $\pm$ 0.3691	304000.0000 $\pm$ 92245.3251	50.0%	175/286
Q-learning ( $\rho = 0.05, \epsilon = 0.25$ )	9.12259 $\pm$ 0.0575	2554175.1000 $\pm$ 11350.4274	0.0%	0.0876 $\pm$ 0.0438	24976.2000 $\pm$ 13129.7253	nan $\pm$ nan	nan $\pm$ nan	0.0%	182/493
Q-learning ( $\rho = 0.1, \epsilon = 0.25$ )	6.26658 $\pm$ 0.0751	1754951.5000 $\pm$ 5525.9510	0.0%	0.0659 $\pm$ 0.0203	18746.5000 $\pm$ 5830.1253	6.0391 $\pm$ 0.1214	1680000.0000 $\pm$ 20542.6386	90.0%	133/347
Q-learning ( $\rho = 0.3, \epsilon = 0.25$ )	4.68124 $\pm$ 0.0436	1311938.9000 $\pm$ 6453.3623	0.0%	0.0395 $\pm$ 0.0270	11493.1000 $\pm$ 7651.9557	2.0760 $\pm$ 0.0508	588000.0000 $\pm$ 9796.5292	50.0%	139/384
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	5.16458 $\pm$ 0.1555	1417405.1000 $\pm$ 7168.2131	0.0%	0.0511 $\pm$ 0.0289	14874.7000 $\pm$ 7885.9615	1.3483 $\pm$ 0.0346	370285.7143 $\pm$ 7435.7300	70.0%	225/411
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	6.42202 $\pm$ 0.0553	1705747.2000 $\pm$ 6643.0314	0.0%	0.0512 $\pm$ 0.0240	13308.6000 $\pm$ 6114.1130	1.0720 $\pm$ 0.0232	285333.3333 $\pm$ 2867.4418	30.0%	300/1142
Q-learning ( $\rho = 0.05, \epsilon = 0.5$ )	9.91728 $\pm$ 0.2256	2778286.5000 $\pm$ 9967.5108	0.0%	0.1145 $\pm$ 0.0646	32119.6000 $\pm$ 17678.3821	nan $\pm$ nan	nan $\pm$ nan	0.0%	178/292
Q-learning ( $\rho = 0.1, \epsilon = 0.5$ )	7.30858 $\pm$ 0.1876	2075725.3000 $\pm$ 11153.2485	0.0%	0.0898 $\pm$ 0.0339	25407.8000 $\pm$ 9106.8728	7.0158 $\pm$ 0.2400	194883.3333 $\pm$ 32780.3106	60.0%	137/213
Q-learning ( $\rho = 0.3, \epsilon = 0.5$ )	6.49163 $\pm$ 0.0532	1823326.1000 $\pm$ 7409.6966	0.0%	0.0872 $\pm$ 0.0454	24703.0000 $\pm$ 12189.4823	2.5204 $\pm$ 0.0687	707500.0000 $\pm$ 20377.6839	100.0%	156/267
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	7.45448 $\pm$ 0.0760	2091213.2000 $\pm$ 10470.8515	0.0%	0.0646 $\pm$ 0.0176	18269.6000 $\pm$ 4873.1855	2.8333 $\pm$ 1.1576	794800.0000 $\pm$ 319459.7940	100.0%	234/535
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	8.98784 $\pm$ 0.0527	2538064.6000 $\pm$ 10811.5659	0.0%	0.0493 $\pm$ 0.0321	13888.5000 $\pm$ 9257.8329	nan $\pm$ nan	nan $\pm$ nan	0.0%	207/837
Q-learning ( $\rho = 0.05, \epsilon = 0.75$ )	12.58560 $\pm$ 0.1962	3677258.2000 $\pm$ 19942.8683	0.0%	0.1077 $\pm$ 0.1077	32086.5000 $\pm$ 31816.4864	nan $\pm$ nan	nan $\pm$ nan	0.0%	138/222
Q-learning ( $\rho = 0.1, \epsilon = 0.75$ )	10.76710 $\pm$ 0.0524	3205292.3000 $\pm$ 13684.9267	0.0%	0.1102 $\pm$ 0.0643	33290.2000 $\pm$ 19055.9031	8.6584 $\pm$ 0.0000	259000.0000 $\pm$ 0.0000	10.0%	163/221
Q-learning ( $\rho = 0.3, \epsilon = 0.75$ )	11.34678 $\pm$ 0.0784	3362790.6000 $\pm$ 17116.1092	0.0%	0.0491 $\pm$ 0.0325	14922.7000 $\pm$ 10020.2733	nan $\pm$ nan	nan $\pm$ nan	0.0%	175/298
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	13.50038 $\pm$ 0.1144	3983012.0000 $\pm$ 27323.2308	0.0%	0.0730 $\pm$ 0.0457	21650.6000 $\pm$ 13761.3699	nan $\pm$ nan	nan $\pm$ nan	0.0%	207/632
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	16.31277 $\pm$ 0.1463	4831096.4000 $\pm$ 20442.7072	0.0%	0.0582 $\pm$ 0.0399	16911.0000 $\pm$ 10777.8656	nan $\pm$ nan	nan $\pm$ nan	0.0%	264/854
Q-learning ( $\rho = 0.05, \epsilon = 0.9$ )	20.90885 $\pm$ 0.5429	6336251.0000 $\pm$ 24239.5722	0.0%	0.1113 $\pm$ 0.0526	33965.9000 $\pm$ 16003.0625	nan $\pm$ nan	nan $\pm$ nan	0.0%	147/235
Q-learning ( $\rho = 0.1, \epsilon = 0.9$ )	20.46911 $\pm$ 0.3372	6326221.1000 $\pm$ 41608.1798	0.0%	0.1141 $\pm$ 0.0729	34939.6000 $\pm$ 22734.0983	nan $\pm$ nan	nan $\pm$ nan	0.0%	159/209
Q-learning ( $\rho = 0.3, \epsilon = 0.9$ )	21.71623 $\pm$ 0.3406	6846501.8000 $\pm$ 40375.1530	0.0%	0.0475 $\pm$ 0.0496	14880.1000 $\pm$ 15556.2382	nan $\pm$ nan	nan $\pm$ nan	0.0%	140/434
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	23.11406 $\pm$ 0.6283	7398753.6000 $\pm$ 41900.8000	0.0%	0.1019 $\pm$ 0.0624	33511.3000 $\pm$ 20074.2052	nan $\pm$ nan	nan $\pm$ nan	0.0%	297/562
Q-learning ( $\rho = 0.7, \epsilon = 0.9$ )	24.36300 $\pm$ 0.5107	7848426.5000 $\pm$ 25471.5155	0.0%	0.1275 $\pm$ 0.0920	41382.4000 $\pm$ 28358.8550	nan $\pm$ nan	nan $\pm$ nan	0.0%	226/724
Q-learning ( $\rho = 0.05, \epsilon = 1$ )	26.72804 $\pm$ 0.4123	886132.9000 $\pm$ 9711.4018	0.0%	0.1990 $\pm$ 0.1340	66993.1000 $\pm$ 45004.6923	nan $\pm$ nan	nan $\pm$ nan	0.0%	123/263
Q-learning ( $\rho = 0.1, \epsilon = 1$ )	69.8524 $\pm$ 128.7632	886612.7000 $\pm$ 9800.8670	0.0%	0.2255 $\pm$ 0.1916	73164.3000 $\pm$ 61230.6265	nan $\pm$ nan	nan $\pm$ nan	0.0%	136/242
Q-learning ( $\rho = 0.3, \epsilon = 1$ )	25.72737 $\pm$ 0.0487	8871062.7000 $\pm$ 6127.5520	0.0%	0.2037 $\pm$ 0.1301	69453.3000 $\pm$ 42509.9341	nan $\pm$ nan	nan $\pm$ nan	0.0%	182/299
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	25.71404 $\pm$ 0.0721	8871489.3000 $\pm$ 10760.8833	0.0%	0.2117 $\pm$ 0.2134	74120.4000 $\pm$ 73008.9996	nan $\pm$ nan	nan $\pm$ nan	0.0%	204/501
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	25.77000 $\pm$ 0.1262	8866343.3000 $\pm$ 9861.5639	0.0%	0.1483 $\pm$ 0.1428	50913.0000 $\pm$ 49458.5697	nan $\pm$ nan	nan $\pm$ nan	0.0%	249/568
Ayco Value Iteration	1.99475 $\pm$ 0.0118	156468.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	146/191
Value Iteration	1.68654 $\pm$ 0.0290	236882.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	132/230

**Table 17.** Performance of DP methods and Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.7$  (Problem 10).

Algorithm (Problem 10) ( $\gamma = 0.7$ )	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.05, \epsilon = 0$ )	5.96387 $\pm$ 0.1696	1512161.4000 $\pm$ 2670.5442	0.0%	0.0579 $\pm$ 0.0476	15139.5000 $\pm$ 12453.2000	nan $\pm$ nan	nan $\pm$ nan	0.0%	119/500
Q-learning ( $\rho = 0.1, \epsilon = 0$ )	3.78183 $\pm$ 0.0686	93473.8000 $\pm$ 2747.9469	0.0%	0.0444 $\pm$ 0.0276	11433.3000 $\pm$ 7109.2537	3.6423 $\pm$ 0.0671	923375.0000 $\pm$ 15206.0982	80.0%	89/136
Q-learning ( $\rho = 0.3, \epsilon = 0$ )	2.16876 $\pm$ 0.0554	552439.9000 $\pm$ 7708.0647	0.0%	0.0339 $\pm$ 0.0230	8278.7000 $\pm$ 5430.2361	1.1778 $\pm$ 0.0485	299500.0000 $\pm$ 3354.1020	100.0%	103/156
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	1.91131 $\pm$ 0.0531	486992.3000 $\pm$ 9675.5909	0.0%	0.0178 $\pm$ 0.0115	4832.2000 $\pm$ 3207.7101	0.9841 $\pm$ 0.2493	250666.6667 $\pm$ 63712.8977	90.0%	93/135
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	1.88296 $\pm$ 0.0723	479389.7000 $\pm$ 11571.0215	0.0%	0.0170 $\pm$ 0.0097	4054.9000 $\pm$ 2071.4557	0.8854 $\pm$ 0.4787	224555.5556 $\pm$ 118481.2326	90.0%	102/164
Q-learning ( $\rho = 0.05, \epsilon = 0.25$ )	6.01242 $\pm$ 0.0394	1563032.3000 $\pm$ 2398.1777	0.0%	0.0386 $\pm$ 0.0308	9979.9000 $\pm$ 7808.7939	nan $\pm$ nan	nan $\pm$ nan	0.0%	104/722
Q-learning ( $\rho = 0.1, \epsilon = 0.25$ )	4.00974 $\pm$ 0.0267	1035449.3000 $\pm$ 3159.8627	0.0%	0.0272 $\pm$ 0.0302	6948.8000 $\pm$ 7603.1372	3.8178 $\pm$ 0.0049	987000.0000 $\pm$ 10000.0000	20.0%	95/125
Q-learning ( $\rho = 0.3, \epsilon = 0.25$ )	0.2651 $\pm$ 0.0261	692706.6000 $\pm$ 3277.2031	0.0%	0.0261 $\pm$ 0.0161	7338.9000 $\pm$ 4366.2303	1.3116 $\pm$ 0.0418	340500.0000 $\pm$ 11800.4237	100.0%	104/128
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	2.65333 $\pm$ 0.0295	689653.7000 $\pm$ 3702.4262	0.0%	0.0254 $\pm$ 0.0178	6484.2000 $\pm$ 4885.6365	0.8223 $\pm$ 0.0439	213100.0000 $\pm$ 11885.9621	100.0%	97/197
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	2.95973 $\pm$ 0.0557	77139.4000 $\pm$ 3641.6586	0.0%	0.0272 $\pm$ 0.0171	7083.4000 $\pm$ 4201.6833	0.6182 $\pm$ 0.0196	155666.6667 $\pm$ 1490.7120	60.0%	148/296
Q-learning ( $\rho = 0.05, \epsilon = 0.5$ )	6.51836 $\pm$ 0.0778	1718152.0000 $\pm$ 4763.9567	0.0%	0.0763 $\pm$ 0.0425	20883.3000 $\pm$ 11174.4661	nan $\pm$ nan	nan $\pm$ nan	0.0%	92/184
Q-learning ( $\rho = 0.1, \epsilon = 0.5$ )	4.91298 $\pm$ 0.0634	1245892.1000 $\pm$ 5218.7915	0.0%	0.0364 $\pm$ 0.0220	9807.3000 $\pm$ 5512.0979	4.6469 $\pm$ 0.0420	1190250.0000 $\pm$ 10825.3175	40.0%	85/127
Q-learning ( $\rho = 0.3, \epsilon = 0.5$ )	3.14506 $\pm$ 0.0729	965302.3000 $\pm$ 2655.3866	0.0%	0.0380 $\pm$ 0.0186	10192.3000 $\pm$ 4855.7640	1.7846 $\pm$ 0.2819	460200.0000 $\pm$ 65978.4813	100.0%	100/142
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	3.86460 $\pm$ 0.0834	1001219.4000 $\pm$ 2588.9391	0.0%	0.0234 $\pm$ 0.0165	6967.5000 $\pm$ 4365.0193	0.9733 $\pm$ 0.0682	255600.0000 $\pm$ 18413.0389	100.0%	99/179
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	4.36818 $\pm$ 0.0934	1129015.9000 $\pm$ 5753.2040	0.0%	0.0216 $\pm$ 0.0094	5799.7000 $\pm$ 2554.9131	0.7853 $\pm$ 0.0964	205800.0000 $\pm$ 25646.8322	100.0%	143/428
Q-learning ( $\rho = 0.05, \epsilon = 0.75$ )	8.75981 $\pm$ 0.1488	2329030.8000 $\pm$ 5585.0317	0.0%	0.1138 $\pm$ 0.0793	31149.9000 $\pm$ 22031.2859	nan $\pm$ nan	nan $\pm$ nan	0.0%	79/119
Q-learning ( $\rho = 0.1, \epsilon = 0.75$ )	7.49800 $\pm$ 0.1085	1989297.4000 $\pm$ 7094.2379	0.0%	0.0618 $\pm$ 0.0334	16473.3000 $\pm$ 8929.8288	5.7761 $\pm$ 0.5224	1562600.0000 $\pm$ 133329.2447	40.0%	92/147
Q-learning ( $\rho = 0.3, \epsilon = 0.75$ )	6.88314 $\pm$ 0.1305	1831675.0000 $\pm$ 6392.2100	0.0%	0.0517 $\pm$ 0.0259	13712.3000 $\pm$ 7438.6561	3.8212 $\pm$ 1.7151	1024000.0000 $\pm$ 465013.1181	50.0%	101/121
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	7.16563 $\pm$ 0.1659	1942609.1000 $\pm$ 8138.6615	0.0%	0.0394 $\pm$ 0.0228	10599.0000 $\pm$ 5948.5857	nan $\pm$ nan	nan $\pm$ nan	0.0%	118/197
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	8.00057 $\pm$ 0.0740	217869.9000 $\pm$ 6554.5616	0.0%	0.0390 $\pm$ 0.0243	10646.1000 $\pm$ 6928.9696	nan $\pm$ nan	nan $\pm$ nan	0.0%	144/436
Q-learning ( $\rho = 0.05, \epsilon = 0.9$ )	15.85728 $\pm$ 0.1505	4443227.4000 $\pm$ 29971.9655	0.0%	0.0858 $\pm$ 0.0549	24843.2000 $\pm$ 16237.4335	11.2460 $\pm$ 0.0000	3143000.0000 $\pm$ 0.0000	10.0%	84/143
Q-learning ( $\rho = 0.1, \epsilon = 0.9$ )	15.7183								

**Table 18.** Performance of DP methods and Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.9$  (Problem 10).

Algorithm (Problem 10)	Run time ( $\gamma = 0.9$ ) (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.05, \epsilon = 0$ )	4.25422 $\pm$ 0.1022	1074244.4000 $\pm$ 443.2784	0.0%	0.0286 $\pm$ 0.0176	7168.8000 $\pm$ 4539.2967	nan $\pm$ nan	nan $\pm$ nan	0.0%	84/29037
Q-learning ( $\rho = 0.1, \epsilon = 0$ )	2.47363 $\pm$ 0.0346	661308.6000 $\pm$ 562.9441	0.0%	0.0142 $\pm$ 0.0190	4155.8000 $\pm$ 5454.1404	2.4213 $\pm$ 0.0263	645666.6667 $\pm$ 2867.4418	60.0%	67/78
Q-learning ( $\rho = 0.3, \epsilon = 0$ )	1.38754 $\pm$ 0.0357	372630.7000 $\pm$ 1890.6645	0.0%	0.0098 $\pm$ 0.0083	2771.1000 $\pm$ 2129.7503	0.7891 $\pm$ 0.0331	213500.0000 $\pm$ 8958.2364	100.0%	69/84
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	1.20380 $\pm$ 0.0140	321507.2000 $\pm$ 5933.9071	0.0%	0.0154 $\pm$ 0.0062	4374.7000 $\pm$ 1932.6496	0.5243 $\pm$ 0.0804	138200.0000 $\pm$ 17803.3705	100.0%	73/91
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	1.15708 $\pm$ 0.0324	309478.0000 $\pm$ 5366.3737	0.0%	0.0082 $\pm$ 0.0025	1862.9000 $\pm$ 916.8512	0.5458 $\pm$ 0.3031	143000.0000 $\pm$ 75529.2438	60.0%	71/89
Q-learning ( $\rho = 0.05, \epsilon = 0.25$ )	4.09429 $\pm$ 0.0360	1118923.6000 $\pm$ 1164.5843	0.0%	0.0295 $\pm$ 0.0161	8258.0000 $\pm$ 4771.8050	nan $\pm$ nan	nan $\pm$ nan	0.0%	73/831
Q-learning ( $\rho = 0.1, \epsilon = 0.25$ )	2.69927 $\pm$ 0.0388	728539.7000 $\pm$ 866.2722	0.0%	0.0220 $\pm$ 0.0132	5969.8000 $\pm$ 3317.4900	2.5393 $\pm$ 0.0587	706800.0000 $\pm$ 18882.7964	50.0%	66/84
Q-learning ( $\rho = 0.3, \epsilon = 0.25$ )	1.62501 $\pm$ 0.0252	460307.9000 $\pm$ 870.4154	0.0%	0.0201 $\pm$ 0.0095	3577.4000 $\pm$ 2530.6853	1.0746 $\pm$ 0.0955	305857.1429 $\pm$ 28973.0068	70.0%	67/88
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	1.49631 $\pm$ 0.0315	418317.9000 $\pm$ 1674.5988	0.0%	0.0158 $\pm$ 0.0071	4227.6000 $\pm$ 1757.0224	0.5465 $\pm$ 0.0697	153400.0000 $\pm$ 20140.5864	100.0%	67/84
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	1.52844 $\pm$ 0.0172	432750.2000 $\pm$ 2054.6122	0.0%	0.0118 $\pm$ 0.0071	3287.1000 $\pm$ 2216.3088	0.4230 $\pm$ 0.0674	120500.0000 $\pm$ 20323.6316	100.0%	78/400
Q-learning ( $\rho = 0.05, \epsilon = 0.5$ )	4.28771 $\pm$ 0.0479	1236475.8000 $\pm$ 2899.7580	0.0%	0.0324 $\pm$ 0.0213	9622.1000 $\pm$ 6321.6981	nan $\pm$ nan	nan $\pm$ nan	0.0%	70/388
Q-learning ( $\rho = 0.1, \epsilon = 0.5$ )	3.07881 $\pm$ 0.0272	880892.2000 $\pm$ 1814.2900	0.0%	0.0281 $\pm$ 0.0191	8646.2000 $\pm$ 5451.9187	2.8153 $\pm$ 0.0983	813375.0000 $\pm$ 33540.7867	80.0%	66/86
Q-learning ( $\rho = 0.3, \epsilon = 0.5$ )	2.23190 $\pm$ 0.0341	658070.8000 $\pm$ 2453.3467	0.0%	0.0189 $\pm$ 0.0137	5419.8000 $\pm$ 3722.3724	1.1581 $\pm$ 0.2314	343000.0000 $\pm$ 69415.9307	70.0%	68/82
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	2.10181 $\pm$ 0.0232	622804.0000 $\pm$ 1562.8830	0.0%	0.0187 $\pm$ 0.0084	5537.9000 $\pm$ 2506.8575	0.6816 $\pm$ 0.1490	202400.0000 $\pm$ 43518.2720	100.0%	70/135
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	2.15394 $\pm$ 0.0292	636542.0000 $\pm$ 1429.2440	0.0%	0.0108 $\pm$ 0.0085	3284.2000 $\pm$ 2427.8488	0.4365 $\pm$ 0.0299	129000.0000 $\pm$ 8763.5609	100.0%	73/269
Q-learning ( $\rho = 0.05, \epsilon = 0.75$ )	5.65393 $\pm$ 0.0683	1703830.6000 $\pm$ 6119.9470	0.0%	0.0545 $\pm$ 0.0373	16082.8000 $\pm$ 11936.7665	nan $\pm$ nan	nan $\pm$ nan	0.0%	69/76
Q-learning ( $\rho = 0.1, \epsilon = 0.75$ )	4.78082 $\pm$ 0.0340	1446867.9000 $\pm$ 4909.8018	0.0%	0.0502 $\pm$ 0.0218	15544.9000 $\pm$ 6677.8183	3.5025 $\pm$ 0.5400	1062250.0000 $\pm$ 158316.7316	80.0%	65/77
Q-learning ( $\rho = 0.3, \epsilon = 0.75$ )	4.27364 $\pm$ 0.0704	1287621.8000 $\pm$ 5580.9577	0.0%	0.0283 $\pm$ 0.0179	8946.7000 $\pm$ 5479.3997	2.2220 $\pm$ 0.6976	666777.7778 $\pm$ 208290.8105	90.0%	69/83
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	4.17856 $\pm$ 0.0413	1270217.3000 $\pm$ 8410.6708	0.0%	0.0226 $\pm$ 0.0117	6767.0000 $\pm$ 3345.5285	1.3636 $\pm$ 0.5554	415000.0000 $\pm$ 171860.9903	100.0%	66/152
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	4.28299 $\pm$ 0.0702	1297171.0000 $\pm$ 7627.5490	0.0%	0.0290 $\pm$ 0.0137	8349.1000 $\pm$ 4065.5056	1.1803 $\pm$ 0.4889	357700.0000 $\pm$ 144956.5707	100.0%	76/148
Q-learning ( $\rho = 0.05, \epsilon = 0.9$ )	10.95564 $\pm$ 0.1662	3385292.5000 $\pm$ 21731.9319	0.0%	0.0418 $\pm$ 0.0216	13515.3000 $\pm$ 7494.2332	8.4737 $\pm$ 1.1036	2621111.1111 $\pm$ 328660.8029	90.0%	75/86
Q-learning ( $\rho = 0.1, \epsilon = 0.9$ )	10.54643 $\pm$ 0.0798	327752.3000 $\pm$ 18161.0016	0.0%	0.0419 $\pm$ 0.0162	12685.3000 $\pm$ 4823.0110	5.4089 $\pm$ 1.4687	1682625.0000 $\pm$ 448392.3888	80.0%	66/79
Q-learning ( $\rho = 0.3, \epsilon = 0.9$ )	10.37067 $\pm$ 0.0760	3219725.5000 $\pm$ 18550.0356	0.0%	0.0292 $\pm$ 0.0231	9303.1000 $\pm$ 7032.5092	3.2454 $\pm$ 1.8469	1038285.7143 $\pm$ 569998.4123	70.0%	68/82
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	10.75918 $\pm$ 0.2029	3227290.3000 $\pm$ 21805.9856	0.0%	0.0315 $\pm$ 0.0183	9817.8000 $\pm$ 5485.1292	4.5028 $\pm$ 2.5028	1378833.3333 $\pm$ 796713.1263	60.0%	69/111
Q-learning ( $\rho = 0.7, \epsilon = 0.9$ )	10.81855 $\pm$ 0.2442	3299951.7000 $\pm$ 19867.0822	0.0%	0.0512 $\pm$ 0.0290	15617.3000 $\pm$ 8658.5057	2.2075 $\pm$ 0.5192	685500.0000 $\pm$ 165500.0000	20.0%	71/213
Q-learning ( $\rho = 0.05, \epsilon = 1$ )	28.41448 $\pm$ 0.5732	8751202.5000 $\pm$ 9411.5658	0.0%	0.0975 $\pm$ 0.0623	30976.8000 $\pm$ 19431.4598	26.8243 $\pm$ 0.4043	8315000.0000 $\pm$ 268000.0000	20.0%	72/83
Q-learning ( $\rho = 0.1, \epsilon = 1$ )	28.09445 $\pm$ 0.3850	8746357.1000 $\pm$ 17683.6414	0.0%	0.1330 $\pm$ 0.1227	42867.7000 $\pm$ 26577.3722	21.1432 $\pm$ 4.3765	1613375.0000 $\pm$ 1345175.1687	80.0%	70/82
Q-learning ( $\rho = 0.3, \epsilon = 1$ )	28.15691 $\pm$ 0.4088	8735736.8000 $\pm$ 16185.1551	0.0%	0.1008 $\pm$ 0.0978	31550.3000 $\pm$ 30688.4362	12.3954 $\pm$ 5.6479	3848300.0000 $\pm$ 172038.4580	100.0%	66/84
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	30.57344 $\pm$ 0.6988	8749544.5000 $\pm$ 12582.7589	0.0%	0.0891 $\pm$ 0.1044	25791.4000 $\pm$ 31442.1964	7.8988 $\pm$ 4.4143	2273800.0000 $\pm$ 1290161.6023	100.0%	71/181
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	30.27275 $\pm$ 0.3097	8748822.0000 $\pm$ 10130.2903	0.0%	0.0439 $\pm$ 0.0194	13493.7000 $\pm$ 6125.6220	16.7331 $\pm$ 9.8434	4789000.0000 $\pm$ 2805490.0891	50.0%	68/86
Async Value Iteration	0.29688 $\pm$ 0.0101	547362.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	67/87
Value Iteration	0.48309 $\pm$ 0.0289	553320.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	69/79

**Table 19.** Performance of DP methods and Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.5$  (Problem 11).

Algorithm (Problem 11)	Run time ( $\gamma = 0.5$ ) (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.05, \epsilon = 0$ )	3.05628 $\pm$ 0.0417	931983.1000 $\pm$ 4335.6100	0.0%	0.0105 $\pm$ 0.0047	3205.2000 $\pm$ 1516.4140	nan $\pm$ nan	nan $\pm$ nan	0.0%	38/238
Q-learning ( $\rho = 0.1, \epsilon = 0$ )	1.59090 $\pm$ 0.0282	60860.1000 $\pm$ 2705.2238	0.0%	0.0089 $\pm$ 0.0055	2631.0000 $\pm$ 1653.4711	1.7480 $\pm$ 0.0448	53700.0000 $\pm$ 8414.8678	100.0%	39/306
Q-learning ( $\rho = 0.3, \epsilon = 0$ )	1.19195 $\pm$ 0.0204	363139.1000 $\pm$ 8142.2723	0.0%	0.0080 $\pm$ 0.0030	2522.5000 $\pm$ 971.9032	0.6873 $\pm$ 0.1315	210000.0000 $\pm$ 40281.5094	100.0%	43/76
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	1.05041 $\pm$ 0.0230	318803.0000 $\pm$ 6969.1476	0.0%	0.0075 $\pm$ 0.0036	2453.6000 $\pm$ 1079.3368	0.4852 $\pm$ 0.1418	148222.2222 $\pm$ 44681.0369	90.0%	47/109
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	1.03261 $\pm$ 0.0488	316623.4000 $\pm$ 14922.6060	0.0%	0.0090 $\pm$ 0.0054	2880.2000 $\pm$ 1719.5827	0.3469 $\pm$ 0.1628	109125.0000 $\pm$ 51933.7017	80.0%	59/126
Q-learning ( $\rho = 0.05, \epsilon = 0.25$ )	3.09059 $\pm$ 0.0334	902644.0000 $\pm$ 4848.7360	0.0%	0.0114 $\pm$ 0.0049	3574.5000 $\pm$ 1554.6038	nan $\pm$ nan	nan $\pm$ nan	0.0%	40/258
Q-learning ( $\rho = 0.1, \epsilon = 0.25$ )	2.12340 $\pm$ 0.0382	660166.5000 $\pm$ 2695.4388	0.0%	0.0123 $\pm$ 0.0036	3998.8000 $\pm$ 1317.5341	1.8310 $\pm$ 0.0649	570000.0000 $\pm$ 14906.3745	100.0%	48/95
Q-learning ( $\rho = 0.3, \epsilon = 0.25$ )	1.54959 $\pm$ 0.0322	481027.4000 $\pm$ 3625.3863	0.0%	0.0108 $\pm$ 0.0063	3528.0000 $\pm$ 2148.9767	0.7006 $\pm$ 0.0974	218900.0000 $\pm$ 30632.6623	100.0%	38/88
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	1.67900 $\pm$ 0.0304	514453.0000 $\pm$ 4841.6983	0.0%	0.0083 $\pm$ 0.0029	2706.9000 $\pm$ 968.5790	0.7044 $\pm$ 0.1412	216571.4286 $\pm$ 122715.5330	70.0%	63/310
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	2.02010 $\pm$ 0.0134	618068.0000 $\pm$ 4699.3731	0.0%	0.0132 $\pm$ 0.0067	3988.8000 $\pm$ 2111.5785	0.3730 $\pm$ 0.1942	116500.0000 $\pm$ 60911.8215	100.0%	102/279
Q-learning ( $\rho = 0.05, \epsilon = 0.5$ )	3.31060 $\pm$ 0.0454	1051958.0000 $\pm$ 7819.2191	0.0%	0.0102 $\pm$ 0.0050	3123.4000 $\pm$ 1505.8675	nan $\pm$ nan	nan $\pm$ nan	0.0%	51/146
Q-learning ( $\rho = 0.1, \epsilon = 0.5$ )	2.47391 $\pm$ 0.0344	789987.2000 $\pm$ 5454.8317	0.0%	0.0098 $\pm$ 0.0069	3081.4000 $\pm$ 1978.9308	1.9975 $\pm$ 0.0654	631888.8889 $\pm$ 17291.2595	90.0%	35/77
Q-learning ( $\rho = 0.3, \epsilon = 0.5$ )	2.10075 $\pm$ 0.0147	667489.6000 $\pm$ 2071.0397	0.0%	0.0146 $\pm$ 0.0062	4812.3000 $\pm$ 1918.8502	0.6891 $\pm$ 0.0249	221800.0000 $\pm$ 7717.5126	100.0%	38/109
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	2.37681 $\pm$ 0.0257	759175.7000 $\pm$ 4896.8396	0.0%	0.0118 $\pm$ 0.0060	3729.3000 $\pm$ 1942.0390	0.5349 $\pm$ 0.1143	174300.0000 $\pm$ 38946.2450	100.0%	66/301
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	2.92918 $\pm$ 0.0494	926039.8000 $\pm$ 10533.7296	0.0%	0.0124 $\pm$ 0.0065	4034.1000 $\pm$ 2156.0668	1.0520 $\pm$ 0.4108	331285.7143 $\pm$ 130550.1702	70.0%	81/275
Q-learning ( $\rho = 0.05, \epsilon = 0.75$ )	4.27226 $\pm$ 0.0507	1389634.9000 $\pm$ 11219.3949	0.0%	0.0064 $\pm$ 0.0032	2182.1000 $\pm$ 1130.1038	nan $\pm$ nan	nan $\pm$ nan	0.0%	48/90
Q-learning ( $\rho = 0.1, \epsilon = 0.75$ )	3.70340 $\pm$ 0.0442	1214265.8000 $\pm$ 1779.8259	0.0%	0.0149 $\pm$ 0.0127	3009.4000 $\pm$ 4065.1467	2.6810 $\pm$ 0.3665	881222.2222 $\pm$ 119554.5229	90.0%	32/102
Q-learning ( $\rho = 0.3, \epsilon = 0.75$ )	3.82740 $\pm$ 0.0634	1242230.3000 $\pm$ 13564.5511	0.0%	0.0088 $\pm$ 0.0052	2930.9000 $\pm$ 1760.8187	2.1078 $\pm$ 0.8438	668857.1429 $\pm$ 277170.1018	70.0%	41/166
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	4.48964 $\pm$ 0.0547	1462640.9000 $\pm$ 15725.0141	0.0%	0.0069 $\pm$ 0.0031	2415.4000 $\pm$ 1131.9916	nan $\pm$ nan	nan $\pm$ nan	0.0%	70/317
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	5.41929 $\pm$ 0.0774	1760862.4000 $\pm$ 12003.8286	0.0%	0.0074 $\pm$ 0.0041	2355.2000 $\pm$				

**Table 20.** Performance of DP methods and Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.7$  (Problem 11).

Algorithm (Problem 11) ( $\gamma = 0.7$ )	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.05, \epsilon = 0$ )	1.85072 $\pm$ 0.0295	571188.4000 $\pm$ 1736.1736	0.0%	0.0075 $\pm$ 0.0034	2390.2000 $\pm$ 1008.7788	nan $\pm$ nan	nan $\pm$ nan	0.0%	29/285
Q-learning ( $\rho = 0.1, \epsilon = 0$ )	1.14824 $\pm$ 0.0237	356490.7000 $\pm$ 1679.6158	0.0%	0.0078 $\pm$ 0.0033	2359.4000 $\pm$ 903.0394	0.9849 $\pm$ 0.0242	30580.0000 $\pm$ 7839.3717	100.0%	31/41
Q-learning ( $\rho = 0.3, \epsilon = 0$ )	0.65108 $\pm$ 0.0161	202669.4000 $\pm$ 8396.0761	0.0%	0.0069 $\pm$ 0.0023	2246.9000 $\pm$ 733.8597	0.3335 $\pm$ 0.0299	104000.0000 $\pm$ 9121.4034	100.0%	29/42
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	0.58352 $\pm$ 0.0212	176486.2000 $\pm$ 5260.8166	0.0%	0.0078 $\pm$ 0.0041	2234.4000 $\pm$ 986.1301	0.2652 $\pm$ 0.0907	81600.0000 $\pm$ 25865.8075	100.0%	31/62
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	0.55968 $\pm$ 0.0267	172331.0000 $\pm$ 5244.1676	0.0%	0.0081 $\pm$ 0.0032	2443.8000 $\pm$ 1067.0102	0.2965 $\pm$ 0.1341	91000.0000 $\pm$ 42092.7547	100.0%	28/52
Q-learning ( $\rho = 0.05, \epsilon = 0.25$ )	1.87500 $\pm$ 0.0236	590310.7000 $\pm$ 2177.8359	0.0%	0.0095 $\pm$ 0.0053	3295.1000 $\pm$ 2025.5409	nan $\pm$ nan	nan $\pm$ nan	0.0%	28/65
Q-learning ( $\rho = 0.1, \epsilon = 0.25$ )	1.24466 $\pm$ 0.0400	389796.7000 $\pm$ 1779.5030	0.0%	0.0091 $\pm$ 0.0033	2694.3000 $\pm$ 919.2288	1.0451 $\pm$ 0.0761	127400.0000 $\pm$ 15588.4573	100.0%	25/57
Q-learning ( $\rho = 0.3, \epsilon = 0.25$ )	0.80251 $\pm$ 0.0194	292618.3000 $\pm$ 1626.5426	0.0%	0.0068 $\pm$ 0.0034	2190.8000 $\pm$ 1129.2834	0.3748 $\pm$ 0.0392	119300.0000 $\pm$ 12720.4560	100.0%	29/118
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	0.78585 $\pm$ 0.0135	249014.0000 $\pm$ 3344.5959	0.0%	0.0067 $\pm$ 0.0030	2190.0000 $\pm$ 799.4007	0.2309 $\pm$ 0.0311	75900.0000 $\pm$ 10520.9315	100.0%	29/75
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	0.88708 $\pm$ 0.0174	275623.1000 $\pm$ 1277.1430	0.0%	0.0069 $\pm$ 0.0042	2172.8000 $\pm$ 1219.5270	0.2573 $\pm$ 0.1515	80100.0000 $\pm$ 46601.3948	100.0%	37/192
Q-learning ( $\rho = 0.05, \epsilon = 0.5$ )	2.03455 $\pm$ 0.0377	65294.6000 $\pm$ 2854.5637	0.0%	0.0078 $\pm$ 0.0031	2561.8000 $\pm$ 1047.6895	nan $\pm$ nan	nan $\pm$ nan	0.0%	29/122
Q-learning ( $\rho = 0.1, \epsilon = 0.5$ )	1.46130 $\pm$ 0.0131	472283.0000 $\pm$ 953.4786	0.0%	0.0099 $\pm$ 0.0064	3090.8000 $\pm$ 1844.6777	1.1263 $\pm$ 0.0681	365800.0000 $\pm$ 24770.1433	100.0%	31/44
Q-learning ( $\rho = 0.3, \epsilon = 0.5$ )	1.09838 $\pm$ 0.0206	354386.5000 $\pm$ 1866.4678	0.0%	0.0087 $\pm$ 0.0035	2771.8000 $\pm$ 1181.1545	0.2983 $\pm$ 0.0440	130500.0000 $\pm$ 13522.2040	100.0%	30/46
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	1.13065 $\pm$ 0.0174	361634.5000 $\pm$ 1266.2115	0.0%	0.0076 $\pm$ 0.0041	2618.4000 $\pm$ 1386.6625	0.2841 $\pm$ 0.0807	92900.0000 $\pm$ 24230.5674	100.0%	36/142
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	1.25506 $\pm$ 0.0138	40482.1000 $\pm$ 2022.1663	0.0%	0.0064 $\pm$ 0.0029	1990.8000 $\pm$ 1007.2874	0.2065 $\pm$ 0.0271	67600.0000 $\pm$ 9624.9675	100.0%	38/368
Q-learning ( $\rho = 0.05, \epsilon = 0.75$ )	2.69296 $\pm$ 0.0191	892406.4000 $\pm$ 5942.4291	0.0%	0.0112 $\pm$ 0.0073	3719.5000 $\pm$ 2462.0602	2.4583 $\pm$ 0.1202	81544.4444 $\pm$ 41131.5265	90.0%	27/57
Q-learning ( $\rho = 0.1, \epsilon = 0.75$ )	2.25077 $\pm$ 0.0233	751469.6000 $\pm$ 3122.7089	0.0%	0.0077 $\pm$ 0.0029	2562.0000 $\pm$ 1019.0408	1.2703 $\pm$ 0.1051	425400.0000 $\pm$ 36540.9359	100.0%	28/44
Q-learning ( $\rho = 0.3, \epsilon = 0.75$ )	2.03944 $\pm$ 0.0189	679102.1000 $\pm$ 4756.2187	0.0%	0.0140 $\pm$ 0.0067	4566.2000 $\pm$ 2354.3572	0.4792 $\pm$ 0.0430	164200.0000 $\pm$ 15548.6334	100.0%	26/49
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	2.17193 $\pm$ 0.0318	715946.6000 $\pm$ 6384.1643	0.0%	0.0097 $\pm$ 0.0062	3138.9000 $\pm$ 1924.8360	0.4323 $\pm$ 0.1421	141800.0000 $\pm$ 47509.5780	100.0%	33/63
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	2.42896 $\pm$ 0.0384	801925.5000 $\pm$ 8217.3459	0.0%	0.0094 $\pm$ 0.0053	3023.2000 $\pm$ 1541.6154	0.4701 $\pm$ 0.1764	158300.0000 $\pm$ 58027.6658	100.0%	35/100
Q-learning ( $\rho = 0.05, \epsilon = 0.9$ )	4.95338 $\pm$ 0.0688	1663924.7000 $\pm$ 19298.1902	0.0%	0.0065 $\pm$ 0.0052	1969.6000 $\pm$ 1381.8728	2.9590 $\pm$ 0.1871	100470.0000 $\pm$ 61796.7573	100.0%	31/55
Q-learning ( $\rho = 0.1, \epsilon = 0.9$ )	4.80960 $\pm$ 0.0639	1626720.0000 $\pm$ 21414.2769	0.0%	0.0060 $\pm$ 0.0050	2086.5000 $\pm$ 1628.0246	1.5324 $\pm$ 0.1175	520800.0000 $\pm$ 40703.3168	100.0%	27/44
Q-learning ( $\rho = 0.3, \epsilon = 0.9$ )	4.86495 $\pm$ 0.0493	1651366.2000 $\pm$ 15328.2049	0.0%	0.0057 $\pm$ 0.0036	1932.0000 $\pm$ 1066.7638	0.6609 $\pm$ 0.1607	228100.0000 $\pm$ 57118.2108	100.0%	27/58
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	5.22074 $\pm$ 0.1232	175077.3000 $\pm$ 14806.5850	0.0%	0.0106 $\pm$ 0.0045	3021.4000 $\pm$ 1571.2568	0.8291 $\pm$ 0.3267	279900.0000 $\pm$ 110134.8719	100.0%	32/81
Q-learning ( $\rho = 0.7, \epsilon = 0.9$ )	6.55502 $\pm$ 0.0808	1912059.2000 $\pm$ 18695.1183	0.0%	0.0048 $\pm$ 0.0029	1601.6000 $\pm$ 1027.8246	2.2342 $\pm$ 0.0000	759000.0000 $\pm$ 0.0000	10.0%	33/88
Q-learning ( $\rho = 0.1, \epsilon = 1$ )	15.41285 $\pm$ 0.0935	531809.5000 $\pm$ 38522.8975	0.0%	0.0071 $\pm$ 0.0047	2445.2000 $\pm$ 1689.7547	1.1668 $\pm$ 1.4009	2127100.0000 $\pm$ 48526.4050	100.0%	27/51
Q-learning ( $\rho = 0.3, \epsilon = 1$ )	15.39910 $\pm$ 0.1935	5310918.5000 $\pm$ 52614.4491	0.0%	0.0065 $\pm$ 0.0060	2345.2000 $\pm$ 2314.2856	3.3551 $\pm$ 0.5765	1156000.0000 $\pm$ 191039.2630	100.0%	25/46
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	15.47807 $\pm$ 0.1695	531695.5000 $\pm$ 5207.1087	0.0%	0.0061 $\pm$ 0.0043	2302.0000 $\pm$ 1415.0763	1.1796 $\pm$ 0.0522	54000.0000 $\pm$ 141861.6523	90.0%	26/52
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	15.47259 $\pm$ 0.2803	5334627.2000 $\pm$ 64805.3006	0.0%	0.0066 $\pm$ 0.0032	2044.2000 $\pm$ 1023.9160	0.5088 $\pm$ 0.3793	1761285.7143 $\pm$ 1225453.5853	70.0%	26/97
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	15.40955 $\pm$ 0.1712	5329222.6000 $\pm$ 49331.8510	0.0%	0.0094 $\pm$ 0.0073	3404.0000 $\pm$ 2618.4177	nan $\pm$ nan	nan $\pm$ nan	0.0%	32/133
Asynic Value Iteration	0.39380 $\pm$ 0.0078	582080.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	31/48
Value Iteration	0.61098 $\pm$ 0.0218	866272.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	29/41

**Table 21.** Performance of DP methods and Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.9$  (Problem 11).

Algorithm (Problem 11) ( $\gamma = 0.9$ )	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.05, \epsilon = 0$ )	1.27863 $\pm$ 0.0220	402517.4000 $\pm$ 1531.6856	0.0%	0.0063 $\pm$ 0.0020	1962.7000 $\pm$ 609.9456	1.2686 $\pm$ 0.0234	399500.0000 $\pm$ 3278.7193	40.0%	25/622
Q-learning ( $\rho = 0.1, \epsilon = 0$ )	0.78346 $\pm$ 0.0149	244012.8000 $\pm$ 1006.6520	0.0%	0.0056 $\pm$ 0.0013	1728.4000 $\pm$ 349.4402	0.6537 $\pm$ 0.0220	204400.0000 $\pm$ 5986.6518	100.0%	23/32
Q-learning ( $\rho = 0.3, \epsilon = 0$ )	0.42620 $\pm$ 0.0100	154923.5000 $\pm$ 1267.5796	0.0%	0.0053 $\pm$ 0.0012	1692.2000 $\pm$ 430.9220	0.2215 $\pm$ 0.0144	70400.0000 $\pm$ 3411.7444	100.0%	22/38
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	0.37807 $\pm$ 0.0201	115781.0000 $\pm$ 3152.5928	0.0%	0.0059 $\pm$ 0.0017	1798.7000 $\pm$ 4119.0761	1.1796 $\pm$ 0.0522	54000.0000 $\pm$ 141861.6523	90.0%	26/52
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	0.35158 $\pm$ 0.0103	109529.8000 $\pm$ 2347.4768	0.0%	0.0076 $\pm$ 0.0061	1675.4000 $\pm$ 436.1360	0.1594 $\pm$ 0.0886	49800.0000 $\pm$ 27128.5827	100.0%	22/31
Q-learning ( $\rho = 0.05, \epsilon = 0.25$ )	1.31733 $\pm$ 0.0158	421869.2000 $\pm$ 1067.8852	0.0%	0.0072 $\pm$ 0.0021	2341.6000 $\pm$ 601.9774	1.3034 $\pm$ 0.0097	418000.0000 $\pm$ 4301.1626	40.0%	23/29
Q-learning ( $\rho = 0.1, \epsilon = 0.25$ )	0.84795 $\pm$ 0.0137	273211.9000 $\pm$ 747.9059	0.0%	0.0060 $\pm$ 0.0037	1993.6000 $\pm$ 949.9605	0.6631 $\pm$ 0.0242	213800.0000 $\pm$ 8518.2158	100.0%	22/30
Q-learning ( $\rho = 0.3, \epsilon = 0.25$ )	0.52191 $\pm$ 0.0226	168453.1000 $\pm$ 603.8706	0.0%	0.0065 $\pm$ 0.0023	1990.0000 $\pm$ 605.5198	0.2362 $\pm$ 0.0223	70400.0000 $\pm$ 6843.9755	100.0%	22/33
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	0.47103 $\pm$ 0.0143	150304.9000 $\pm$ 799.3068	0.0%	0.0048 $\pm$ 0.0018	1637.8000 $\pm$ 494.4348	0.1729 $\pm$ 0.0533	55300.0000 $\pm$ 10344.5638	100.0%	22/30
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	0.48917 $\pm$ 0.0140	135594.8000 $\pm$ 1154.1807	0.0%	0.0065 $\pm$ 0.0008	1936.3000 $\pm$ 387.2782	0.1570 $\pm$ 0.0340	49100.0000 $\pm$ 10074.2245	100.0%	23/161
Q-learning ( $\rho = 0.05, \epsilon = 0.5$ )	1.45333 $\pm$ 0.0133	472961.3000 $\pm$ 1386.4713	0.0%	0.0080 $\pm$ 0.0028	2730.3000 $\pm$ 808.2796	1.3476 $\pm$ 0.0289	443600.0000 $\pm$ 8879.1892	100.0%	22/29
Q-learning ( $\rho = 0.1, \epsilon = 0.5$ )	1.02841 $\pm$ 0.0125	357065.0000 $\pm$ 816.3183	0.0%	0.0088 $\pm$ 0.0023	3119.2000 $\pm$ 882.9344	0.0871 $\pm$ 0.0228	22900.0000 $\pm$ 5692.0968	100.0%	23/33
Q-learning ( $\rho = 0.3, \epsilon = 0.5$ )	0.74511 $\pm$ 0.0119	242345.7000 $\pm$ 950.0345	0.0%	0.0082 $\pm$ 0.0024	2677.6000 $\pm$ 785.5479	0.2627 $\pm$ 0.0294	88800.0000 $\pm$ 7547.1849	100.0%	23/30
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	0.68277 $\pm$ 0.0106	224632.0000 $\pm$ 818.2979	0.0%	0.0076 $\pm$ 0.0025	2580.5000 $\pm$ 770.9068	0.1686 $\pm$ 0.0207	56900.0000 $\pm$ 7183.3140	100.0%	22/37
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	0.65065 $\pm$ 0.0094	227732.3000 $\pm$ 1397.4116	0.0%	0.0064 $\pm$ 0.0017	2167.3000 $\pm$ 607.9712	0.1483 $\pm$ 0.0341	49300.0000 $\pm$ 10982.2584	100.0%	24/30
Q-learning ( $\rho = 0.05, \epsilon = 0.75$ )	1.93422 $\pm$ 0.0226	654523.1000 $\pm$ 2364.1268	0.0%	0.0069 $\pm$ 0.0044	3357.4000 $\pm$ 1379.9961	1.5492 $\pm$ 0.0578	524500.0000 $\pm$ 13843.7712	100.0%	22/36
Q-learning ( $\rho = 0.1, \epsilon = 0.75$ )	1.63222 $\pm$ 0.0211	549482.2000 $\pm$ 2786.6671	0.0%	0.0081 $\pm$ 0.0021	2814.4000 $\pm$ 908.8171	0.7917 $\pm$ 0.0435	268300.0000 $\pm$ 13520.7248	100.0%	22/29
Q-learning ( $\rho = 0.3, \epsilon = 0.75$ )	1.42428 $\pm$ 0.0276	476723.7000 $\pm$ 3385.8993	0.0%	0.0080 $\pm$ 0.0045	2788.3000 $\pm$ 1691.7654	0.2968 $\pm$ 0.0298	102700.0000 $\pm$ 9839.2073	100.0%	22/32
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	1.58091 $\pm$ 0.0135	465512.0000 $\pm$ 3544.0541	0.0%	0.0079 $\pm$ 0.0020	2541.1000 $\pm$ 597.5407	0.2098 $\pm$ 0.0215	70100.0000 $\pm$ 6122.8078	100.0%	23/29
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	1.41781 $\pm$ 0.0196	475307.6000 $\pm$ 2213.8200	0.						

**Table 23.** Performance of DP methods and Q-learning as we change the learning rate  $\rho$  across Problems 1, 10, 11 with  $\gamma = 0.7$  and  $\epsilon$ -decay.

Problem	Algorithm ( $\gamma = 0.7$ )	Runtime (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Problem 1	Q-learning ( $\rho = 0.1, \epsilon$ -decay)	2.16270 $\pm$ 0.0419	658836.4000 $\pm$ 9598.6608	0.0%	0.0067 $\pm$ 0.0069	2638.3000 $\pm$ 2274.6604	2.1215 $\pm$ 0.0848	644400.0000 $\pm$ 16740.3704	50.0%	36/72
	Q-learning ( $\rho = 0.5, \epsilon$ -decay)	2.33761 $\pm$ 0.0476	71369.3000 $\pm$ 10360.6532	0.0%	0.0057 $\pm$ 0.0046	1664.8000 $\pm$ 1389.9136	2.2215 $\pm$ 0.0736	682400.0000 $\pm$ 15007.9970	100.0%	43/89
	Q-learning ( $\rho = 0.7, \epsilon$ -decay)	2.54275 $\pm$ 0.0339	779184.0000 $\pm$ 9794.5332	0.0%	0.0087 $\pm$ 0.0050	2668.1000 $\pm$ 1413.2517	2.3245 $\pm$ 0.0358	716100.0000 $\pm$ 12856.5159	100.0%	40/964
Problem 10	Q-learning ( $\rho = 0.1, \epsilon$ -decay)	5.95362 $\pm$ 0.2101	1740023.6000 $\pm$ 12322.4373	0.0%	0.0666 $\pm$ 0.0791	20799.5000 $\pm$ 25673.8536	8.6562 $\pm$ 0.3261	1657250.0000 $\pm$ 72912.1903	40.0%	90/117
	Q-learning ( $\rho = 0.5, \epsilon$ -decay)	6.25790 $\pm$ 0.1271	189067.9000 $\pm$ 19252.7588	0.0%	0.1419 $\pm$ 0.1046	44964.1000 $\pm$ 33965.3956	5.2121 $\pm$ 0.2097	1500800.0000 $\pm$ 48186.7202	100.0%	106/221
	Q-learning ( $\rho = 0.7, \epsilon$ -decay)	6.85639 $\pm$ 0.1001	208075.9000 $\pm$ 11179.6210	0.0%	0.1271 $\pm$ 0.0869	38283.9000 $\pm$ 25201.6927	5.6864 $\pm$ 0.2092	172100.0000 $\pm$ 52065.0378	100.0%	143/292
Problem 11	Q-learning ( $\rho = 0.1, \epsilon$ -decay)	2.80780 $\pm$ 0.0611	696267.3000 $\pm$ 15229.0235	0.0%	0.0065 $\pm$ 0.0044	1979.8000 $\pm$ 1376.0974	1.7137 $\pm$ 0.0921	525600.0000 $\pm$ 27372.2487	100.0%	24/52
	Q-learning ( $\rho = 0.5, \epsilon$ -decay)	2.50699 $\pm$ 0.0438	746330.8000 $\pm$ 15052.4150	0.0%	0.0069 $\pm$ 0.0062	2160.8000 $\pm$ 1684.5450	1.2700 $\pm$ 0.4061	387800.0000 $\pm$ 121489.7527	100.0%	34/60
	Q-learning ( $\rho = 0.7, \epsilon$ -decay)	2.87144 $\pm$ 0.0884	839811.5000 $\pm$ 12281.4056	0.0%	0.0101 $\pm$ 0.0097	2863.2000 $\pm$ 3077.0052	1.9816 $\pm$ 0.3205	583300.0000 $\pm$ 88221.3693	100.0%	35/195

**Table 24.** Performance of DP methods and Q-learning as we change the learning rate  $\rho$  across Problems 1, 10, 11 with  $\gamma = 0.9$  and  $\epsilon$ -decay.

Problem	Algorithm ( $\gamma = 0.9$ )	Runtime (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Problem 1	Q-learning ( $\rho = 0.1, \epsilon$ -decay)	1.66952 $\pm$ 0.0451	508233.7000 $\pm$ 6252.7432	0.0%	0.0067 $\pm$ 0.0072	1956.8000 $\pm$ 2326.8822	1.5912 $\pm$ 0.0565	488800.0000 $\pm$ 14161.9208	50.0%	28/42
	Q-learning ( $\rho = 0.5, \epsilon$ -decay)	1.66440 $\pm$ 0.0528	519443.7000 $\pm$ 11570.6701	0.0%	0.0061 $\pm$ 0.0057	2065.3000 $\pm$ 1886.0563	1.6242 $\pm$ 0.0713	500000.0000 $\pm$ 19869.5747	50.0%	29/37
	Q-learning ( $\rho = 0.7, \epsilon$ -decay)	1.69829 $\pm$ 0.0224	524086.2000 $\pm$ 7062.3417	0.0%	0.0076 $\pm$ 0.0045	2102.8000 $\pm$ 1467.6715	1.5281 $\pm$ 0.1464	475000.0000 $\pm$ 43000.0000	100.0%	29/169
Problem 10	Q-learning ( $\rho = 0.1, \epsilon$ -decay)	4.45455 $\pm$ 0.1874	1352699.6000 $\pm$ 12188.7971	0.0%	0.0929 $\pm$ 0.0452	29464.5000 $\pm$ 14158.1076	4.0246 $\pm$ 0.3538	1227500.0000 $\pm$ 69655.2223	100.0%	71/83
	Q-learning ( $\rho = 0.5, \epsilon$ -decay)	4.57553 $\pm$ 0.1357	1359892.0000 $\pm$ 8836.6398	0.0%	0.1469 $\pm$ 0.0801	45129.3000 $\pm$ 26001.8712	3.8665 $\pm$ 0.3422	808800.0000 $\pm$ 278771.5193	100.0%	73/267
	Q-learning ( $\rho = 0.7, \epsilon$ -decay)	4.80031 $\pm$ 0.2351	1404558.4000 $\pm$ 16347.0547	0.0%	0.1590 $\pm$ 0.0824	48680.1000 $\pm$ 25518.9627	3.6637 $\pm$ 0.4287	1080400.0000 $\pm$ 98097.0948	100.0%	91/176
Problem 11	Q-learning ( $\rho = 0.1, \epsilon$ -decay)	1.71151 $\pm$ 0.0329	527798.8000 $\pm$ 5334.1851	0.0%	0.0109 $\pm$ 0.0088	3629.6000 $\pm$ 2917.8036	1.1859 $\pm$ 0.0716	369800.0000 $\pm$ 20183.1613	100.0%	22/28
	Q-learning ( $\rho = 0.5, \epsilon$ -decay)	1.72770 $\pm$ 0.0276	529884.5000 $\pm$ 7678.3650	0.0%	0.0054 $\pm$ 0.0055	1691.9000 $\pm$ 1648.9380	0.4592 $\pm$ 0.0760	144000.0000 $\pm$ 22834.1849	100.0%	22/36
	Q-learning ( $\rho = 0.7, \epsilon$ -decay)	1.80766 $\pm$ 0.0332	540734.0000 $\pm$ 9597.4643	0.0%	0.0054 $\pm$ 0.0031	1609.6000 $\pm$ 808.9669	0.3562 $\pm$ 0.0092	113800.0000 $\pm$ 22964.3202	100.0%	22/261

**Table 25.** Performance of DP methods and Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.5$  (Problem 8).

Algorithm (Problem 8) ( $\gamma = 0.5$ )	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.1, \epsilon = 0$ )	0.13995 $\pm$ 0.0056	44668.1000 $\pm$ 954.4866	0.0%	0.0010 $\pm$ 0.0000	60.2000 $\pm$ 24.3343	0.0371 $\pm$ 0.0107	12000.0000 $\pm$ 3521.3634	100.0%	7/36
Q-learning ( $\rho = 0.1, \epsilon = 0.25$ )	0.17386 $\pm$ 0.0069	56086.3000 $\pm$ 461.1481	0.0%	0.0010 $\pm$ 0.0000	71.0000 $\pm$ 46.9766	0.0349 $\pm$ 0.0166	11800.0000 $\pm$ 5741.0800	100.0%	6/21
Q-learning ( $\rho = 0.1, \epsilon = 0.5$ )	0.22990 $\pm$ 0.0060	74276.3000 $\pm$ 451.2032	0.0%	nan $\pm$ nan	60.9000 $\pm$ 43.7762	0.0318 $\pm$ 0.0152	10800.0000 $\pm$ 6686.1468	100.0%	8/21
Q-learning ( $\rho = 0.1, \epsilon = 0.75$ )	0.31864 $\pm$ 0.0099	107960.5000 $\pm$ 1346.5570	0.0%	0.0011 $\pm$ 0.0001	58.3000 $\pm$ 37.5379	0.0471 $\pm$ 0.0234	16100.0000 $\pm$ 7542.5460	100.0%	8/40
Q-learning ( $\rho = 0.1, \epsilon = 0.9$ )	0.41847 $\pm$ 0.0080	144128.9000 $\pm$ 1953.6720	0.0%	0.0010 $\pm$ 0.0000	69.9000 $\pm$ 37.6708	0.0474 $\pm$ 0.0300	16300.0000 $\pm$ 9908.6696	100.0%	7/67
Q-learning ( $\rho = 0.1, \epsilon = 1$ )	0.53000 $\pm$ 0.0215	181779.4000 $\pm$ 1735.9936	0.0%	0.0010 $\pm$ 0.0000	72.5000 $\pm$ 30.3883	0.0548 $\pm$ 0.0155	19000.0000 $\pm$ 5549.7748	100.0%	6/31
Q-learning ( $\rho = \frac{1}{n(n+1)}, \epsilon = 0$ )	0.14912 $\pm$ 0.0099	43173.7000 $\pm$ 579.2908	0.0%	0.0010 $\pm$ 0.0000	70.7000 $\pm$ 31.8467	0.0861 $\pm$ 0.0151	24857.1429 $\pm$ 5890.1569	70.0%	9/27
Q-learning ( $\rho = \frac{1}{n(n+1)}, \epsilon = 0.25$ )	0.17867 $\pm$ 0.0057	52892.4000 $\pm$ 679.6827	0.0%	0.0010 $\pm$ 0.0000	71.1000 $\pm$ 53.0367	0.1180 $\pm$ 0.0425	34400.0000 $\pm$ 12354.7562	50.0%	8/28
Q-learning ( $\rho = \frac{1}{n(n+1)}, \epsilon = 0.5$ )	0.22891 $\pm$ 0.0133	69617.6000 $\pm$ 3901.1876	10.0%	0.0010 $\pm$ 0.0000	89.0000 $\pm$ 42.8416	0.1179 $\pm$ 0.0322	36666.6667 $\pm$ 10811.5165	60.0%	6/35
Q-learning ( $\rho = \frac{1}{n(n+1)}, \epsilon = 0.75$ )	0.33659 $\pm$ 0.0047	104463.0000 $\pm$ 1254.3633	0.0%	0.0013 $\pm$ 0.0005	94.2000 $\pm$ 50.7756	0.1568 $\pm$ 0.0468	48333.3333 $\pm$ 14290.6311	60.0%	12/24
Q-learning ( $\rho = \frac{1}{n(n+1)}, \epsilon = 0.9$ )	0.42532 $\pm$ 0.0468	133795.0000 $\pm$ 16046.5053	20.0%	0.0010 $\pm$ 0.0000	80.4000 $\pm$ 59.3013	0.2706 $\pm$ 0.0882	54875.0000 $\pm$ 27342.4464	80.0%	7/28
Q-learning ( $\rho = \frac{1}{n(n+1)}, \epsilon = 1$ )	0.56828 $\pm$ 0.0132	181020.5000 $\pm$ 2239.2016	0.0%	0.0010 $\pm$ 0.0000	44.0000 $\pm$ 22.7420	0.3087 $\pm$ 0.0925	98500.0000 $\pm$ 28860.2957	60.0%	9/24

**Table 26.** Performance of DP methods and Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.7$  (Problem 8).

Algorithm (Problem 8) ( $\gamma = 0.7$ )	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.1, \epsilon = 0$ )	0.08366 $\pm$ 0.0022	26825.4000 $\pm$ 129.0102	0.0%	0.0010 $\pm$ 0.0000	55.1000 $\pm$ 10.0941	0.0163 $\pm$ 0.0089	5200.0000 $\pm$ 2749.5454	100.0%	6/11
Q-learning ( $\rho = 0.1, \epsilon = 0.25$ )	0.10487 $\pm$ 0.0029	34645.2000 $\pm$ 337.1219	0.0%	nan $\pm$ nan	54.4000 $\pm$ 17.3966	0.0167 $\pm$ 0.0029	5400.0000 $\pm$ 1113.5529	100.0%	7/17
Q-learning ( $\rho = 0.1, \epsilon = 0.5$ )	0.14780 $\pm$ 0.0060	48376.7000 $\pm$ 401.5768	0.0%	0.0010 $\pm$ 0.0000	75.5000 $\pm$ 36.2039	0.0250 $\pm$ 0.0260	8400.0000 $\pm$ 8357.0330	100.0%	6/12
Q-learning ( $\rho = 0.1, \epsilon = 0.75$ )	0.22654 $\pm$ 0.0099	76800.8000 $\pm$ 716.2207	0.0%	0.0014 $\pm$ 0.0005	82.0000 $\pm$ 45.4626	0.0261 $\pm$ 0.0201	8000.0000 $\pm$ 6315.8531	100.0%	6/12
Q-learning ( $\rho = 0.1, \epsilon = 0.9$ )	0.32994 $\pm$ 0.0105	113612.1000 $\pm$ 1430.1477	0.0%	0.0010 $\pm$ 0.0000	29.8000 $\pm$ 14.6820	0.0237 $\pm$ 0.0070	8000.0000 $\pm$ 2607.6810	100.0%	6/11
Q-learning ( $\rho = 0.1, \epsilon = 1$ )	0.46176 $\pm$ 0.0121	162542.8000 $\pm$ 1633.1799	0.0%	0.0010 $\pm$ 0.0000	43.9000 $\pm$ 32.7275	0.0309 $\pm$ 0.0200	10900.0000 $\pm$ 6655.0733	100.0%	6/13
Q-learning ( $\rho = \frac{1}{n(n+1)}, \epsilon = 0$ )	0.09052 $\pm$ 0.0108	24763.2000 $\pm$ 2091.1724	20.0%	0.0010 $\pm$ 0.0000	54.4000 $\pm$ 15.1142	0.0280 $\pm$ 0.0176	7800.0000 $\pm$ 4770.7442	100.0%	6/15
Q-learning ( $\rho = \frac{1}{n(n+1)}, \epsilon = 0.25$ )	0.09668 $\pm$ 0.0242	28498.0000 $\pm$ 6484.9181	60.0%	0.0010 $\pm$ 0.0000	62.4000 $\pm$ 15.3571	0.0242 $\pm$ 0.0090	7000.0000 $\pm$ 2607.6810	100.0%	7/17
Q-learning ( $\rho = \frac{1}{n(n+1)}, \epsilon = 0.5$ )	0.12390 $\pm$ 0.0372	36596.0000 $\pm$ 9908.6227	60.0%	0.0010 $\pm$ 0.0000	64.5000 $\pm$ 23.4659	0.0254 $\pm$ 0.0073	7400.0000 $\pm$ 2244.9944	100.0%	6/15
Q-learning ( $\rho = \frac{1}{n(n+1)}, \epsilon = 0.75$ )	0.12229 $\pm$ 0.0471	38215.6000 $\pm$ 14606.2351	100.0%	0.0012 $\pm$ 0.0000	53.3000 $\pm$ 18.9681	0.0362 $\pm$ 0.0181	11400.0000 $\pm$ 5642.6944	100.0%	6/14
Q-learning ( $\rho = \frac{1}{n(n+1)}, \epsilon = 0.9$ )	0.22085 $\pm$ 0.0954	69579.4000 $\pm$ 29671.2468	90.0%	0.0010 $\pm$ 0.0000	62.9000 $\pm$ 36.2642	0.0435 $\pm$ 0.0146	13800.0000 $\pm$ 4955.8047	100.0%	6/11
Q-learning ( $\rho = \frac{1}{n(n+1)}, \epsilon = 1$ )	0.31037 $\pm$ 0.1316	99554.0000 $\pm$ 41351.2545	80.0%	0.0009 $\pm$ 0.0000	41.7000 $\pm$ 27.8138	0.0644 $\pm$ 0.0255	20900.0000 $\pm$ 9027.1812	100.0%	6/12

**Table 27.** Performance of DP methods and Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.9$  (Problem 8).

Algorithm (Problem 8) ( $\gamma = 0.9$ )	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.1, \epsilon = 0$ )	0.06127 $\pm$ 0.0056	18524.3000 $\pm$ 54.7669	0.0%	0.0010 $\pm$ 0.0000	48.2000 $\pm$ 5.9967	0.0135 $\pm$ 0.0053	4100.0000 $\pm$ 1757.8396	100.0%	6/8
Q-learning ( $\rho = 0.1, \epsilon = 0.25$ )	0.08170 $\pm$ 0.0057	24880.8000 $\pm$ 140.5125	0.0%	0.0010 $\pm$ 0.0001	54.8000 $\pm$ 11.6944	0.0119 $\pm$ 0.0039	3700.0000 $\pm$ 1268.8578	100.0%	6/8

**Table 28.** Performance of DP methods and Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.5$  (Problem 16).

Algorithm (Problem 16) ( $\gamma = 0.5$ )	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.1, \epsilon = 0$ )	0.12982 $\pm$ 0.0100	41389.8000 $\pm$ 4158.8417	10.0%	0.0010 $\pm$ 0.0000	39.2000 $\pm$ 20.4929	0.0482 $\pm$ 0.0255	16375.0000 $\pm$ 8484.3666	80.0%	6/15
Q-learning ( $\rho = 0.1, \epsilon = 0.25$ )	0.15886 $\pm$ 0.0042	53968.8000 $\pm$ 620.2851	0.0%	0.0010 $\pm$ 0.0000	59.3000 $\pm$ 43.9342	0.0585 $\pm$ 0.0469	19600.0000 $\pm$ 15344.0542	100.0%	5/33
Q-learning ( $\rho = 0.1, \epsilon = 0.5$ )	0.18802 $\pm$ 0.0049	65436.1000 $\pm$ 720.4785	0.0%	nan $\pm$ nan	31.9000 $\pm$ 19.5164	0.0498 $\pm$ 0.0374	17300.0000 $\pm$ 13183.7021	100.0%	5/21
Q-learning ( $\rho = 0.1, \epsilon = 0.75$ )	0.22302 $\pm$ 0.0086	81339.0000 $\pm$ 1646.8644	0.0%	0.0010 $\pm$ 0.0000	28.6000 $\pm$ 12.8590	0.0478 $\pm$ 0.0484	16900.0000 $\pm$ 17397.9884	100.0%	10/56
Q-learning ( $\rho = 0.1, \epsilon = 0.9$ )	0.26826 $\pm$ 0.0107	94253.2000 $\pm$ 1388.8168	0.0%	nan $\pm$ nan	50.2000 $\pm$ 55.3711	0.0349 $\pm$ 0.0296	12100.0000 $\pm$ 10348.4298	100.0%	5/29
Q-learning ( $\rho = 0.1, \epsilon = 1$ )	0.29285 $\pm$ 0.0062	104646.4000 $\pm$ 2076.6652	0.0%	0.0011 $\pm$ 0.0001	30.6000 $\pm$ 15.9762	0.0648 $\pm$ 0.0501	23100.0000 $\pm$ 17026.1563	100.0%	7/28
Q-learning ( $\rho = \frac{1}{\pi^2 \omega}$ ) ( $\epsilon = 0$ )	0.12767 $\pm$ 0.0217	38422.6000 $\pm$ 6241.5515	30.0%	0.0010 $\pm$ 0.0000	48.8000 $\pm$ 37.8042	0.0885 $\pm$ 0.0245	27000.0000 $\pm$ 7483.3148	60.0%	5/36
Q-learning ( $\rho = \frac{1}{\pi^2 \omega}$ ) ( $\epsilon = 0.25$ )	0.13465 $\pm$ 0.0287	42266.5000 $\pm$ 9628.2294	50.0%	nan $\pm$ nan	26.5000 $\pm$ 8.8798	0.0957 $\pm$ 0.0302	29714.2857 $\pm$ 9779.1949	70.0%	5/28
Q-learning ( $\rho = \frac{1}{\pi^2 \omega}$ ) ( $\epsilon = 0.5$ )	0.19607 $\pm$ 0.0106	61402.8000 $\pm$ 3189.0828	10.0%	nan $\pm$ nan	28.1000 $\pm$ 14.8624	0.1384 $\pm$ 0.0486	43750.0000 $\pm$ 15554.3402	80.0%	5/51
Q-learning ( $\rho = \frac{1}{\pi^2 \omega}$ ) ( $\epsilon = 0.75$ )	0.21758 $\pm$ 0.0316	69623.3000 $\pm$ 10517.6685	50.0%	0.0010 $\pm$ 0.0000	44.2000 $\pm$ 31.1634	0.1212 $\pm$ 0.0389	39000.0000 $\pm$ 12247.4487	70.0%	5/17
Q-learning ( $\rho = \frac{1}{\pi^2 \omega}$ ) ( $\epsilon = 0.9$ )	0.25670 $\pm$ 0.0721	80731.8000 $\pm$ 22291.8848	30.0%	nan $\pm$ nan	35.3000 $\pm$ 35.6147	0.1261 $\pm$ 0.0868	38571.4286 $\pm$ 23993.1963	70.0%	7/19
Q-learning ( $\rho = \frac{1}{\pi^2 \omega}$ ) ( $\epsilon = 1$ )	0.27503 $\pm$ 0.0483	90129.6000 $\pm$ 14557.5210	70.0%	0.0010 $\pm$ 0.0000	29.5000 $\pm$ 21.7083	0.1364 $\pm$ 0.0440	44500.0000 $\pm$ 15001.6666	100.0%	5/22

**Table 29.** Performance of DP methods and Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.7$  (Problem 16).

Algorithm (Problem 16) ( $\gamma = 0.7$ )	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.1, \epsilon = 0$ )	0.06361 $\pm$ 0.0184	19818.4000 $\pm$ 5344.0573	20.0%	nan $\pm$ nan	26.6000 $\pm$ 10.6141	0.0817 $\pm$ 0.0182	4500.0000 $\pm$ 5835.2578	100.0%	5/13
Q-learning ( $\rho = 0.1, \epsilon = 0.25$ )	0.08052 $\pm$ 0.0034	27563.9000 $\pm$ 281.9179	0.0%	0.0010 $\pm$ 0.0000	32.8000 $\pm$ 24.2226	0.0162 $\pm$ 0.0105	5600.0000 $\pm$ 3411.7444	100.0%	5/18
Q-learning ( $\rho = 0.1, \epsilon = 0.5$ )	0.10427 $\pm$ 0.0084	36515.4000 $\pm$ 2258.4497	10.0%	0.0010 $\pm$ 0.0000	26.3000 $\pm$ 13.4911	0.0242 $\pm$ 0.0080	8200.0000 $\pm$ 2712.9320	100.0%	5/21
Q-learning ( $\rho = 0.1, \epsilon = 0.75$ )	0.15310 $\pm$ 0.0062	53967.9000 $\pm$ 1433.2967	10.0%	0.0010 $\pm$ 0.0000	22.9000 $\pm$ 19.0916	0.0202 $\pm$ 0.0185	6800.0000 $\pm$ 6209.6699	100.0%	5/13
Q-learning ( $\rho = 0.1, \epsilon = 0.9$ )	0.20158 $\pm$ 0.0093	71726.7000 $\pm$ 857.7608	0.0%	0.0010 $\pm$ 0.0000	25.6000 $\pm$ 11.9348	0.0208 $\pm$ 0.0134	7400.0000 $\pm$ 4800.0000	100.0%	5/16
Q-learning ( $\rho = 0.1, \epsilon = 1$ )	0.24490 $\pm$ 0.0097	87816.0000 $\pm$ 1119.1048	0.0%	0.0010 $\pm$ 0.0000	21.9000 $\pm$ 11.2854	0.0143 $\pm$ 0.0083	5900.0000 $\pm$ 3033.1562	100.0%	5/10
Q-learning ( $\rho = \frac{1}{\pi^2 \omega}$ ) ( $\epsilon = 0$ )	0.02409 $\pm$ 0.0162	7103.7000 $\pm$ 4949.0689	100.0%	0.0010 $\pm$ 0.0000	25.9000 $\pm$ 18.4578	0.0099 $\pm$ 0.0063	2800.0000 $\pm$ 1536.2291	100.0%	5/13
Q-learning ( $\rho = \frac{1}{\pi^2 \omega}$ ) ( $\epsilon = 0.25$ )	0.02524 $\pm$ 0.0117	7805.6000 $\pm$ 3735.9932	100.0%	0.0010 $\pm$ 0.0000	23.6000 $\pm$ 8.3570	0.0068 $\pm$ 0.0031	2000.0000 $\pm$ 894.4272	100.0%	5/20
Q-learning ( $\rho = \frac{1}{\pi^2 \omega}$ ) ( $\epsilon = 0.5$ )	0.02912 $\pm$ 0.0077	9298.7000 $\pm$ 2478.5133	100.0%	0.0008 $\pm$ 0.0002	23.9000 $\pm$ 8.0554	0.0153 $\pm$ 0.0100	4800.0000 $\pm$ 3249.6154	100.0%	5/17
Q-learning ( $\rho = \frac{1}{\pi^2 \omega}$ ) ( $\epsilon = 0.75$ )	0.04472 $\pm$ 0.0277	14314.8000 $\pm$ 8905.4751	100.0%	0.0010 $\pm$ 0.0000	34.6000 $\pm$ 29.5845	0.0170 $\pm$ 0.0066	5400.0000 $\pm$ 2059.1260	100.0%	5/11
Q-learning ( $\rho = \frac{1}{\pi^2 \omega}$ ) ( $\epsilon = 0.9$ )	0.04224 $\pm$ 0.0271	13520.0000 $\pm$ 9318.7356	100.0%	0.0010 $\pm$ 0.0000	24.4000 $\pm$ 19.0536	0.0223 $\pm$ 0.0210	7200.0000 $\pm$ 7263.6079	100.0%	6/14
Q-learning ( $\rho = \frac{1}{\pi^2 \omega}$ ) ( $\epsilon = 1$ )	0.04713 $\pm$ 0.0323	15417.2000 $\pm$ 10797.2421	100.0%	0.0010 $\pm$ 0.0000	31.3000 $\pm$ 28.4817	0.0126 $\pm$ 0.0060	4000.0000 $\pm$ 2190.8902	100.0%	6/16

**Table 30.** Performance of DP methods and Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.9$  (Problem 16).

Algorithm (Problem 16) ( $\gamma = 0.9$ )	Run time (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.1, \epsilon = 0$ )	0.02624 $\pm$ 0.0189	8753.4000 $\pm$ 6058.6937	50.0%	nan $\pm$ nan	16.9000 $\pm$ 4.9890	0.0035 $\pm$ 0.0013	1300.0000 $\pm$ 458.2576	100.0%	5/7
Q-learning ( $\rho = 0.1, \epsilon = 0.25$ )	0.03005 $\pm$ 0.0181	10329.0000 $\pm$ 6178.4042	70.0%	nan $\pm$ nan	22.8000 $\pm$ 11.3384	0.0060 $\pm$ 0.0043	2100.0000 $\pm$ 1577.9734	100.0%	5/9
Q-learning ( $\rho = 0.1, \epsilon = 0.5$ )	0.03391 $\pm$ 0.0151	11905.3000 $\pm$ 5261.5667	100.0%	0.0010 $\pm$ 0.0000	17.1000 $\pm$ 4.4822	0.0073 $\pm$ 0.0057	2600.0000 $\pm$ 2107.1308	100.0%	5/7
Q-learning ( $\rho = 0.1, \epsilon = 0.75$ )	0.04512 $\pm$ 0.0133	15109.6000 $\pm$ 10382.8197	100.0%	0.0010 $\pm$ 0.0000	17.7000 $\pm$ 8.3560	0.0084 $\pm$ 0.0057	2800.0000 $\pm$ 1886.7962	100.0%	5/7
Q-learning ( $\rho = 0.1, \epsilon = 0.9$ )	0.04857 $\pm$ 0.0329	18009.6000 $\pm$ 12123.9792	100.0%	0.0010 $\pm$ 0.0000	20.3000 $\pm$ 10.2864	0.0058 $\pm$ 0.0028	2100.0000 $\pm$ 1044.0307	100.0%	5/7
Q-learning ( $\rho = 0.1, \epsilon = 1$ )	0.04663 $\pm$ 0.0486	16819.9000 $\pm$ 17603.7841	100.0%	0.0007 $\pm$ 0.0003	25.0000 $\pm$ 18.4120	0.0111 $\pm$ 0.0077	3700.0000 $\pm$ 2282.5424	100.0%	5/9
Q-learning ( $\rho = \frac{1}{\pi^2 \omega}$ ) ( $\epsilon = 0$ )	0.00990 $\pm$ 0.0054	3102.0000 $\pm$ 1758.0113	100.0%	0.0011 $\pm$ 0.0000	17.3000 $\pm$ 8.0131	0.0056 $\pm$ 0.0025	1700.0000 $\pm$ 781.0250	100.0%	5/10
Q-learning ( $\rho = \frac{1}{\pi^2 \omega}$ ) ( $\epsilon = 0.25$ )	0.00769 $\pm$ 0.0041	2203.3000 $\pm$ 1165.3702	100.0%	nan $\pm$ nan	20.6000 $\pm$ 7.3103	0.0046 $\pm$ 0.0016	1300.0000 $\pm$ 458.2576	100.0%	5/8
Q-learning ( $\rho = \frac{1}{\pi^2 \omega}$ ) ( $\epsilon = 0.5$ )	0.00522 $\pm$ 0.0028	1603.4000 $\pm$ 800.9541	100.0%	0.0010 $\pm$ 0.0000	19.3000 $\pm$ 6.9289	0.0044 $\pm$ 0.0025	1300.0000 $\pm$ 640.3124	100.0%	5/8
Q-learning ( $\rho = \frac{1}{\pi^2 \omega}$ ) ( $\epsilon = 0.75$ )	0.00918 $\pm$ 0.0044	2808.2000 $\pm$ 1323.9158	100.0%	nan $\pm$ nan	20.5000 $\pm$ 10.0722	0.0064 $\pm$ 0.0021	2000.0000 $\pm$ 894.4272	100.0%	5/7
Q-learning ( $\rho = \frac{1}{\pi^2 \omega}$ ) ( $\epsilon = 0.9$ )	0.01082 $\pm$ 0.0053	3519.4000 $\pm$ 1687.8762	100.0%	0.0010 $\pm$ 0.0000	39.8000 $\pm$ 28.1169	0.0083 $\pm$ 0.0062	2700.0000 $\pm$ 1951.9221	100.0%	5/7
Q-learning ( $\rho = \frac{1}{\pi^2 \omega}$ ) ( $\epsilon = 1$ )	0.01248 $\pm$ 0.0133	3919.2000 $\pm$ 4109.3820	100.0%	nan $\pm$ nan	18.5000 $\pm$ 11.5866	0.0052 $\pm$ 0.0022	1700.0000 $\pm$ 640.3124	100.0%	5/11

**Table 31.** Comparing performance of Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.99$  (Problem 0). Two versions of value iteration are also appended for contrast

Algorithm (Problem 0) ( $\gamma = 0.99$ )	Runtime (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	0.29275 $\pm$ 0.0132	84080.9000 $\pm$ 1846.8434	0.0%	0.0057 $\pm$ 0.0013	1967.5000 $\pm$ 306.3907	0.1255 $\pm$ 0.0115	36400.0000 $\pm$ 3086.6593	100.0%	18/20
Q-learning ( $\rho = 0.1, \epsilon = 0$ )	0.26636 $\pm$ 0.0149	79924.8000 $\pm$ 2140.9895	0.0%	0.0065 $\pm$ 0.0018	1879.2000 $\pm$ 323.0742	0.1033 $\pm$ 0.0160	30250.0000 $\pm$ 4264.6805	80.0%	18/14
Q-learning ( $\rho = 0.9, \epsilon = 0$ )	0.27654 $\pm$ 0.0171	80032.2000 $\pm$ 3053.1595	0.0%	0.0061 $\pm$ 0.0009	1855.3000 $\pm$ 164.8211	0.0892 $\pm$ 0.0526	26500.0000 $\pm$ 15394.8043	80.0%	18/23
Q-learning ( $\rho = 0.99, \epsilon = 0$ )	0.27131 $\pm$ 0.0086	81149.9000 $\pm$ 2000.3920	0.0%	0.0056 $\pm$ 0.0011	1722.4000 $\pm$ 188.7470	0.2524 $\pm$ 0.0000	76000.0000 $\pm$ 0.0000	10.0%	18/24
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	0.36719 $\pm$ 0.0107	111279.2000 $\pm$ 292.4270	0.0%	0.0109 $\pm$ 0.0014	2919.9000 $\pm$ 214.8583	0.1260 $\pm$ 0.0040	37800.0000 $\pm$ 748.3315	100.0%	18/20
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	0.34508 $\pm$ 0.0084	102965.2000 $\pm$ 626.5941	0.0%	0.0092 $\pm$ 0.0010	2662.1000 $\pm$ 307.7028	0.1028 $\pm$ 0.0103	31200.0000 $\pm$ 3370.4599	100.0%	18/21
Q-learning ( $\rho = 0.9, \epsilon = 0.25$ )	0.32791 $\pm$ 0.0084	99398.8000 $\pm$ 820.5098	0.0%	0.0085 $\pm$ 0.0011	2554.2000 $\pm$ 161.4861	0.1068 $\pm$ 0.0389	32800.0000 $\pm$ 12335.3152	100.0%	18/21
Q-learning ( $\rho = 0.99, \epsilon = 0.25$ )	0.32642 $\pm$ 0.0084	99296.9000 $\pm$ 859.8902	0.0%	0.0083 $\pm$ 0.0009	2534.2000 $\pm$ 228.0088	0.2041 $\pm$ 0.0803	63666.6667 $\pm$ 24870.7771	60.0%	18/20
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	0.52353 $\pm$ 0.0127	163961.0000 $\pm$ 616.8108	0.0%	0.0119 $\pm$ 0.0015	3638.8000 $\pm$ 338.8661	0.1414 $\pm$ 0.0087	44600.0000 $\pm$ 3527.0384	100.0%	18/20
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	0.51433 $\pm$ 0.0258	155895.7000 $\pm$ 828.0888	0.0%	0.0121 $\pm$ 0.0020	3510.1000 $\pm$ 294.2127	0.1107 $\pm$ 0.0083	33800.0000 $\pm$ 1400.0000	100.0%	18/20
Q-learning ( $\rho = 0.9, \epsilon = 0.5$ )	0.48894 $\pm$ 0.0179	151876.9000 $\pm$ 636.4334	0.0%	0.0120 $\pm$ 0.0028	3323.0000 $\pm$ 342.3699	0.1198 $\pm$ 0.0408	36700.0000 $\pm$ 12329.2336	100.0%	18/20
Q-learning ( $\rho = 0.99, \epsilon = 0.5$ )	0.48487 $\pm$ 0.0111	152445.2000 $\pm$ 1059.5566	0.0%	0.0110 $\pm$ 0.0017	3386.7000 $\pm$ 394.3341	0.2083 $\pm$ 0.0834	6400		

**Table 32.** Comparing performance of Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.999$  (Problem 0). Two versions of value iteration are also appended for contrast

Algorithm (Problem 0) ( $\gamma = 0.999$ )	Runtime (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	0.27917 $\pm$ 0.0141	81180.4000 $\pm$ 168.4661	0.0%	0.0052 $\pm$ 0.0010	1574.9000 $\pm$ 56.2040	0.1184 $\pm$ 0.0102	34200.0000 $\pm$ 600.0000	100.0%	18/18
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	0.26669 $\pm$ 0.0288	73372.5000 $\pm$ 210.4553	0.0%	0.0032 $\pm$ 0.0023	1657.2000 $\pm$ 103.7630	0.0918 $\pm$ 0.0149	25200.0000 $\pm$ 400.0000	100.0%	18/18
Q-learning ( $\rho = 0.9, \epsilon = 0$ )	0.22992 $\pm$ 0.0057	69749.8000 $\pm$ 575.6870	0.0%	0.0054 $\pm$ 0.0007	1579.3000 $\pm$ 62.5620	0.0684 $\pm$ 0.0039	20400.0000 $\pm$ 800.0000	100.0%	18/20
Q-learning ( $\rho = 0.99, \epsilon = 0$ )	0.23145 $\pm$ 0.0126	70374.2000 $\pm$ 2204.0658	0.0%	0.0050 $\pm$ 0.0006	1554.4000 $\pm$ 90.5044	0.0625 $\pm$ 0.0033	19333.3333 $\pm$ 816.4966	90.0%	18/20
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	0.35672 $\pm$ 0.0059	108971.1000 $\pm$ 529.4137	0.0%	0.0089 $\pm$ 0.0021	2812.9000 $\pm$ 446.2006	0.1235 $\pm$ 0.0045	37100.0000 $\pm$ 300.0000	100.0%	18/19
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	0.33175 $\pm$ 0.0126	102219.0000 $\pm$ 305.3116	0.0%	0.0082 $\pm$ 0.0012	2548.6000 $\pm$ 203.3791	0.0924 $\pm$ 0.0032	27900.0000 $\pm$ 538.5165	100.0%	18/18
Q-learning ( $\rho = 0.9, \epsilon = 0.25$ )	0.31436 $\pm$ 0.0079	96314.3000 $\pm$ 329.4285	0.0%	0.0085 $\pm$ 0.0016	2585.8000 $\pm$ 365.8658	0.0732 $\pm$ 0.0045	22100.0000 $\pm$ 300.0000	100.0%	18/18
Q-learning ( $\rho = 0.99, \epsilon = 0.25$ )	0.31680 $\pm$ 0.0073	96356.5000 $\pm$ 611.5939	0.0%	0.0082 $\pm$ 0.0014	2614.5000 $\pm$ 317.4631	0.0685 $\pm$ 0.0032	21000.0000 $\pm$ 447.2136	100.0%	18/18
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	0.51660 $\pm$ 0.0110	160177.5000 $\pm$ 909.7031	0.0%	0.0107 $\pm$ 0.0018	3508.9000 $\pm$ 342.5149	0.1372 $\pm$ 0.0054	42200.0000 $\pm$ 400.0000	100.0%	18/18
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	0.48789 $\pm$ 0.0084	154458.7000 $\pm$ 1116.4002	0.0%	0.0111 $\pm$ 0.0017	3425.3000 $\pm$ 423.0629	0.1020 $\pm$ 0.0025	32100.0000 $\pm$ 300.0000	100.0%	18/18
Q-learning ( $\rho = 0.9, \epsilon = 0.5$ )	0.48580 $\pm$ 0.0073	153836.3000 $\pm$ 1518.3678	0.0%	0.0114 $\pm$ 0.0016	3474.4000 $\pm$ 328.8663	0.0849 $\pm$ 0.0051	26900.0000 $\pm$ 830.6624	100.0%	18/18
Q-learning ( $\rho = 0.99, \epsilon = 0.5$ )	0.52147 $\pm$ 0.0247	160492.7000 $\pm$ 1837.0062	0.0%	0.0099 $\pm$ 0.0017	3181.8000 $\pm$ 402.3264	0.0799 $\pm$ 0.0025	25400.0000 $\pm$ 489.8979	100.0%	18/19
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	1.04260 $\pm$ 0.0694	341004.1000 $\pm$ 7067.5569	0.0%	0.0151 $\pm$ 0.0049	4830.3000 $\pm$ 1113.3672	0.1628 $\pm$ 0.0157	53300.0000 $\pm$ 1004.9876	100.0%	18/18
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	1.09828 $\pm$ 0.0847	352123.3000 $\pm$ 5513.8133	0.0%	0.0143 $\pm$ 0.0035	4408.8000 $\pm$ 628.1254	0.1307 $\pm$ 0.0103	42100.0000 $\pm$ 943.3981	100.0%	18/20
Q-learning ( $\rho = 0.9, \epsilon = 0.75$ )	1.20328 $\pm$ 0.0317	303504.4000 $\pm$ 8608.0014	0.0%	0.0140 $\pm$ 0.0031	4230.1000 $\pm$ 880.7672	0.1190 $\pm$ 0.0064	36000.0000 $\pm$ 1095.4451	100.0%	18/18
Q-learning ( $\rho = 0.99, \epsilon = 0.75$ )	1.34013 $\pm$ 0.0416	429710.4000 $\pm$ 7278.7790	0.0%	0.0121 $\pm$ 0.0028	3983.6000 $\pm$ 953.3367	0.1077 $\pm$ 0.0071	34500.0000 $\pm$ 1627.8821	100.0%	18/19
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	0.99946 $\pm$ 0.1518	334382.4000 $\pm$ 45421.2774	100.0%	0.0112 $\pm$ 0.0054	3520.9000 $\pm$ 1323.5073	0.2142 $\pm$ 0.0158	70500.0000 $\pm$ 2801.7851	100.0%	18/18
Q-learning ( $\rho = 0.7, \epsilon = 0.9$ )	0.77184 $\pm$ 0.1358	263606.7000 $\pm$ 47668.5157	100.0%	0.0116 $\pm$ 0.0043	3979.0000 $\pm$ 1540.8138	0.1631 $\pm$ 0.0057	54700.0000 $\pm$ 1552.4175	100.0%	18/18
Q-learning ( $\rho = 0.9, \epsilon = 0.9$ )	0.65527 $\pm$ 0.0829	209086.3000 $\pm$ 18341.2354	100.0%	0.0106 $\pm$ 0.0042	3591.6000 $\pm$ 1285.8675	0.1468 $\pm$ 0.0099	48400.0000 $\pm$ 2154.9659	100.0%	18/18
Q-learning ( $\rho = 0.99, \epsilon = 0.9$ )	0.64452 $\pm$ 0.1320	179820.8000 $\pm$ 16648.6672	100.0%	0.0145 $\pm$ 0.0069	4169.9000 $\pm$ 1901.6330	0.1643 $\pm$ 0.0460	46400.0000 $\pm$ 1907.8784	100.0%	18/18
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	0.74025 $\pm$ 0.1282	213183.3000 $\pm$ 14395.8089	100.0%	0.0061 $\pm$ 0.0060	1798.9000 $\pm$ 1729.0229	0.3303 $\pm$ 0.0757	95000.0000 $\pm$ 7771.7437	100.0%	18/18
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	0.60477 $\pm$ 0.1074	167509.4000 $\pm$ 12420.1262	100.0%	0.0112 $\pm$ 0.0110	1994.9000 $\pm$ 1630.5474	0.2904 $\pm$ 0.0727	77600.0000 $\pm$ 7459.2225	100.0%	18/18
Q-learning ( $\rho = 0.9, \epsilon = 1$ )	0.30257 $\pm$ 0.0890	146319.0000 $\pm$ 123531.8339	100.0%	0.0089 $\pm$ 0.0031	1249.3000 $\pm$ 824.5814	0.1913 $\pm$ 0.1033	60222.2222 $\pm$ 32054.7571	100.0%	18/18
Q-learning ( $\rho = 0.99, \epsilon = 1$ )	0.50919 $\pm$ 0.1118	143606.5000 $\pm$ 18301.5489	100.0%	0.0050 $\pm$ 0.0034	1497.0000 $\pm$ 1140.6033	0.2263 $\pm$ 0.0370	64600.0000 $\pm$ 3666.0066	100.0%	18/18
Async Value Iteration	0.12769 $\pm$ 0.0224	177448.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	18/18
Value Iteration	0.16512 $\pm$ 0.0425	229056.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	18/18

**Table 33.** Comparing performance of Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.99$  (Problem 2). Two versions of value iteration are also appended for contrast

Algorithm (Problem 2) ( $\gamma = 0.99$ )	Runtime (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	0.38042 $\pm$ 0.0107	117180.8000 $\pm$ 85.4222	0.0%	0.0032 $\pm$ 0.0009	991.8000 $\pm$ 251.1975	0.1427 $\pm$ 0.0110	44200.0000 $\pm$ 2712.9320	100.0%	30/31
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	0.35280 $\pm$ 0.0100	10759.3000 $\pm$ 1304.7708	0.0%	0.0032 $\pm$ 0.0009	974.8000 $\pm$ 185.1523	0.1083 $\pm$ 0.0107	32700.0000 $\pm$ 2647.6405	100.0%	30/32
Q-learning ( $\rho = 0.9, \epsilon = 0$ )	0.36663 $\pm$ 0.0224	109215.5000 $\pm$ 1517.1100	0.0%	0.0033 $\pm$ 0.0011	999.0000 $\pm$ 274.0759	0.0868 $\pm$ 0.0092	26200.0000 $\pm$ 2521.9040	100.0%	30/308
Q-learning ( $\rho = 0.99, \epsilon = 0$ )	0.35544 $\pm$ 0.0085	109650.5000 $\pm$ 1558.8350	0.0%	0.0027 $\pm$ 0.0015	860.2000 $\pm$ 366.6846	0.1494 $\pm$ 0.0523	4474.2857 $\pm$ 14771.7325	70.0%	30/36
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	0.47809 $\pm$ 0.0085	151503.0000 $\pm$ 254.3887	0.0%	0.0037 $\pm$ 0.0013	1166.5000 $\pm$ 400.4625	0.1958 $\pm$ 0.0295	62200.0000 $\pm$ 8863.4079	100.0%	30/32
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	0.43357 $\pm$ 0.0066	141102.0000 $\pm$ 364.9177	0.0%	0.0037 $\pm$ 0.0009	1249.3000 $\pm$ 224.8514	0.1913 $\pm$ 0.1033	60222.2222 $\pm$ 32054.7571	100.0%	30/34
Q-learning ( $\rho = 0.9, \epsilon = 0.25$ )	0.45130 $\pm$ 0.0140	140627.8000 $\pm$ 412.1582	0.0%	0.0039 $\pm$ 0.0017	1143.6000 $\pm$ 478.0871	0.1857 $\pm$ 0.0937	58500.0000 $\pm$ 30500.0000	20.0%	30/31
Q-learning ( $\rho = 0.99, \epsilon = 0.25$ )	0.44438 $\pm$ 0.0087	139857.5000 $\pm$ 391.0667	0.0%	0.0041 $\pm$ 0.0016	1236.3000 $\pm$ 444.9135	nan $\pm$ nan	nan $\pm$ nan	0.0%	30/33
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	0.69795 $\pm$ 0.0119	224149.9000 $\pm$ 884.7431	0.0%	0.0058 $\pm$ 0.0026	1882.9000 $\pm$ 892.0393	0.2534 $\pm$ 0.0829	81500.0000 $\pm$ 25831.1827	100.0%	30/34
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	0.67533 $\pm$ 0.0114	217890.8000 $\pm$ 1170.2144	0.0%	0.0044 $\pm$ 0.0016	1380.1000 $\pm$ 598.6783	0.1717 $\pm$ 0.1220	102620.0000 $\pm$ 40493.6528	100.0%	30/31
Q-learning ( $\rho = 0.9, \epsilon = 0.5$ )	0.66646 $\pm$ 0.0091	221497.8000 $\pm$ 720.2024	0.0%	0.0067 $\pm$ 0.0031	2142.5000 $\pm$ 874.4046	0.1926 $\pm$ 0.0761	62500.0000 $\pm$ 24500.0000	20.0%	30/32
Q-learning ( $\rho = 0.99, \epsilon = 0.5$ )	0.66310 $\pm$ 0.0108	214407.5000 $\pm$ 781.2234	0.0%	0.0043 $\pm$ 0.0016	1384.1000 $\pm$ 524.9748	nan $\pm$ nan	nan $\pm$ nan	0.0%	30/32
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	1.42015 $\pm$ 0.0129	471036.8000 $\pm$ 3448.4362	0.0%	0.0066 $\pm$ 0.0029	2242.7000 $\pm$ 984.2487	0.6290 $\pm$ 0.3200	208444.4444 $\pm$ 105667.3093	90.0%	30/33
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	1.03779 $\pm$ 0.0162	467047.8000 $\pm$ 2444.8073	0.0%	0.0026 $\pm$ 0.0026	1966.1000 $\pm$ 777.4724	0.7160 $\pm$ 0.1066	241125.0000 $\pm$ 107490.3459	80.0%	30/32
Q-learning ( $\rho = 0.9, \epsilon = 0.75$ )	1.41994 $\pm$ 0.0211	471120.3000 $\pm$ 3976.5778	0.0%	0.0067 $\pm$ 0.0049	2320.0000 $\pm$ 1454.0864	1.0996 $\pm$ 0.0689	362666.6667 $\pm$ 21929.1789	30.0%	30/1773
Q-learning ( $\rho = 0.99, \epsilon = 0.75$ )	1.44096 $\pm$ 0.0122	477468.3000 $\pm$ 1642.4698	0.0%	0.0073 $\pm$ 0.0023	2398.5000 $\pm$ 1099.7910	nan $\pm$ nan	nan $\pm$ nan	0.0%	30/32
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	3.80891 $\pm$ 0.0409	1280614.8000 $\pm$ 11464.5819	0.0%	0.0062 $\pm$ 0.0024	1949.9000 $\pm$ 814.3826	0.6808 $\pm$ 0.4437	228000.0000 $\pm$ 147197.6902	100.0%	30/32
Q-learning ( $\rho = 0.7, \epsilon = 0.9$ )	3.82324 $\pm$ 0.0397	1291548.1000 $\pm$ 10953.6311	0.0%	0.0094 $\pm$ 0.0064	3205.1000 $\pm$ 2062.1897	0.7794 $\pm$ 0.3948	263400.0000 $\pm$ 130880.4190	100.0%	30/31
Q-learning ( $\rho = 0.9, \epsilon = 0.9$ )	3.87834 $\pm$ 0.0649	1309331.5000 $\pm$ 20062.2936	0.0%	0.0067 $\pm$ 0.0023	2035.4000 $\pm$ 814.8687	1.5721 $\pm$ 1.0533	513444.4444 $\pm$ 351328.9409	90.0%	30/32
Q-learning ( $\rho = 0.99, \epsilon = 0.9$ )	4.11874 $\pm$ 0.0570	1381097.7000 $\pm$ 12916.8569	0.0%	0.0072 $\pm$ 0.0035	2372.5000 $\pm$ 1232.2879	nan $\pm$ nan	nan $\pm$ nan	0.0%	30/32
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	15.96155 $\pm$ 0.1005	5434790.2000 $\pm$ 33827.7444	0.0%	0.0072 $\pm$ 0.0052	2469.3000 $\pm$ 1684.1995	1.1400 $\pm$ 0.4229	387400.0000 $\pm$ 142776.1885	100.0%	30/31
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	15.98967 $\pm$ 0.2010	5452153.0000 $\pm$ 49448.4142	0.0%	0.0040 $\pm$ 0.0041	1295.0000 $\pm$ 1388.0005	0.6707 $\pm$ 0.3531	236000.0000 $\pm$ 123698.3095	100.0%	30/31
Q-learning ( $\rho = 0.9, \epsilon = 1$ )	15.98565 $\pm$ 0.1534	5436481.1000 $\pm$ 48745.8184	0.0%	0.0055 $\pm$ 0.0041	1862.2000 $\pm$ 1310.0016	2.9607 $\pm$ 2.2130	100088.8889 $\pm$ 751642.7113	10.0%	30/32
Q-learning ( $\rho = 0.99, \epsilon = 1$ )	15.98029 $\pm$ 0.1735	5427919.6000 $\pm$ 67890.6968	0.0%	0.0052 $\pm$ 0.0022	1667.9000 $\pm$ 755.2251	2.6790 $\pm$ 0.0000	897000.0000 $\pm$ 0.0000	90.0%	30/32
Async Value Iteration	0.16512 $\pm$ 0.0425	103150.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	30/32
Value Iteration	0.13184 $\pm$ 0.0026	186184.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	30/31

**Table 34.** Comparing performance of Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.999$  (Problem 2). Two versions of value iteration are also appended for contrast



**Table 38.** Comparing performance of Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.999$  (Problem 4). Two versions of value iteration are also appended for contrast

Algorithm (Problem 4) ( $\gamma = 0.999$ )	Runtime (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	0.40388 $\pm$ 0.0179	121020.0000 $\pm$ 137.4831	0.0%	0.0060 $\pm$ 0.0012	1810.0000 $\pm$ 291.3884	0.1476 $\pm$ 0.0064	44400.0000 $\pm$ 663.3250	100.0%	32/32
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	0.37363 $\pm$ 0.0109	113288.0000 $\pm$ 71.1632	0.0%	0.0053 $\pm$ 0.0013	1862.0000 $\pm$ 325.1220	0.1055 $\pm$ 0.0050	32300.0000 $\pm$ 0.0000	100.0%	32/130
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	0.36194 $\pm$ 0.0089	109375.8000 $\pm$ 190.3107	0.0%	0.0057 $\pm$ 0.0014	1732.0000 $\pm$ 261.4330	0.0793 $\pm$ 0.0060	24600.0000 $\pm$ 489.8979	100.0%	32/32
Q-learning ( $\rho = 0.99, \epsilon = 0$ )	0.35485 $\pm$ 0.0136	108461.6000 $\pm$ 218.4038	0.0%	0.0066 $\pm$ 0.0016	1841.5000 $\pm$ 342.7907	0.0734 $\pm$ 0.0037	22000.0000 $\pm$ 0.0000	100.0%	32/32
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	0.49504 $\pm$ 0.0162	154548.3000 $\pm$ 448.5671	0.0%	0.0069 $\pm$ 0.0017	2180.9000 $\pm$ 425.9602	0.1603 $\pm$ 0.0099	49700.0000 $\pm$ 640.3124	100.0%	32/32
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	0.46988 $\pm$ 0.0135	146897.7000 $\pm$ 574.0019	0.0%	0.0068 $\pm$ 0.0014	1991.7000 $\pm$ 321.9975	0.1155 $\pm$ 0.0089	36800.0000 $\pm$ 447.2196	100.0%	32/32
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	0.46353 $\pm$ 0.0139	143821.4000 $\pm$ 481.2052	0.0%	0.0056 $\pm$ 0.0010	1825.6000 $\pm$ 398.4219	0.0894 $\pm$ 0.0042	28400.0000 $\pm$ 489.8979	100.0%	32/32
Q-learning ( $\rho = 0.99, \epsilon = 0.25$ )	0.46397 $\pm$ 0.0113	144073.9000 $\pm$ 520.8655	0.0%	0.0057 $\pm$ 0.0018	1761.4000 $\pm$ 357.4555	0.0812 $\pm$ 0.0044	25700.0000 $\pm$ 458.2576	100.0%	32/32
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	0.69296 $\pm$ 0.0182	223784.9000 $\pm$ 1279.7536	0.0%	0.0084 $\pm$ 0.0032	2616.5000 $\pm$ 762.0247	0.1786 $\pm$ 0.0086	57800.0000 $\pm$ 748.3315	100.0%	32/32
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	0.69119 $\pm$ 0.0262	219532.7000 $\pm$ 1334.0115	0.0%	0.0074 $\pm$ 0.0026	2464.8000 $\pm$ 823.2115	0.1344 $\pm$ 0.0063	42900.0000 $\pm$ 538.5155	100.0%	32/34
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	0.69944 $\pm$ 0.0101	219236.7000 $\pm$ 1706.7261	0.0%	0.0076 $\pm$ 0.0014	2390.3000 $\pm$ 470.6778	0.1104 $\pm$ 0.0050	34600.0000 $\pm$ 800.0000	100.0%	32/32
Q-learning ( $\rho = 0.99, \epsilon = 0.5$ )	0.69905 $\pm$ 0.0148	221584.3000 $\pm$ 627.8771	0.0%	0.0072 $\pm$ 0.0015	2298.4000 $\pm$ 471.4453	0.0957 $\pm$ 0.0058	31300.0000 $\pm$ 781.0250	100.0%	32/32
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	1.39298 $\pm$ 0.0279	451896.4000 $\pm$ 5107.7125	0.0%	0.0086 $\pm$ 0.0022	2908.8000 $\pm$ 904.4703	0.2212 $\pm$ 0.0092	72700.0000 $\pm$ 1268.8578	100.0%	32/32
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	1.41149 $\pm$ 0.0325	464625.9000 $\pm$ 2045.6726	0.0%	0.0107 $\pm$ 0.0037	3462.7000 $\pm$ 985.1840	0.1709 $\pm$ 0.0109	56300.0000 $\pm$ 1486.0069	100.0%	32/32
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	1.48087 $\pm$ 0.0261	479659.4000 $\pm$ 4032.2816	0.0%	0.0114 $\pm$ 0.0038	3664.8000 $\pm$ 1090.2903	0.1417 $\pm$ 0.0055	46900.0000 $\pm$ 700.0000	100.0%	32/32
Q-learning ( $\rho = 0.99, \epsilon = 0.75$ )	1.51294 $\pm$ 0.0371	491389.0000 $\pm$ 3876.1084	0.0%	0.0110 $\pm$ 0.0033	3501.6000 $\pm$ 1062.4862	0.1314 $\pm$ 0.0069	43200.0000 $\pm$ 1939.0719	100.0%	32/32
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	3.99713 $\pm$ 0.2393	1319832.5000 $\pm$ 79436.8863	20.0%	0.0119 $\pm$ 0.0084	4102.4000 $\pm$ 2792.8856	0.2828 $\pm$ 0.0127	95600.0000 $\pm$ 3039.7368	100.0%	32/32
Q-learning ( $\rho = 0.7, \epsilon = 0.9$ )	3.91978 $\pm$ 0.0811	1199225.9000 $\pm$ 218482.7231	70.0%	0.0116 $\pm$ 0.0055	3824.0000 $\pm$ 1580.0028	0.2137 $\pm$ 0.0136	72500.0000 $\pm$ 3170.1735	100.0%	32/34
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	3.52656 $\pm$ 0.8051	1176102.4000 $\pm$ 263042.2913	70.0%	0.0154 $\pm$ 0.0063	5068.8000 $\pm$ 1926.3449	0.1860 $\pm$ 0.0125	64100.0000 $\pm$ 3590.2646	100.0%	32/32
Q-learning ( $\rho = 0.99, \epsilon = 0.9$ )	1.76258 $\pm$ 0.6522	575884.7000 $\pm$ 212264.7682	100.0%	0.0102 $\pm$ 0.0035	3479.4000 $\pm$ 1287.8182	0.1818 $\pm$ 0.0161	60500.0000 $\pm$ 2291.2878	100.0%	32/32
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	1.20305 $\pm$ 0.1918	405112.7000 $\pm$ 60689.9822	100.0%	0.0051 $\pm$ 0.0044	1613.7000 $\pm$ 1544.9670	0.6304 $\pm$ 0.0550	213900.0000 $\pm$ 19684.7657	100.0%	32/32
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	1.02961 $\pm$ 0.1794	344522.6000 $\pm$ 62566.4433	100.0%	0.0085 $\pm$ 0.0076	2666.1000 $\pm$ 2884.2685	0.4529 $\pm$ 0.0710	153300.0000 $\pm$ 21900.1305	100.0%	32/32
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	0.96460 $\pm$ 0.2231	324083.5000 $\pm$ 71469.0675	100.0%	0.0064 $\pm$ 0.0065	2042.1000 $\pm$ 1891.6086	0.4776 $\pm$ 0.0802	160000.0000 $\pm$ 26164.8619	100.0%	32/32
Q-learning ( $\rho = 0.99, \epsilon = 1$ )	0.96659 $\pm$ 0.2904	319446.0000 $\pm$ 96412.0417	100.0%	0.0110 $\pm$ 0.0095	3823.2000 $\pm$ 3389.0056	0.3749 $\pm$ 0.0847	123500.0000 $\pm$ 24422.3259	100.0%	32/32
Async Value Iteration	0.09035 $\pm$ 0.0026	125114.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	32/32
Value Iteration	0.11383 $\pm$ 0.0066	154396.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	32/32

**Table 39.** Comparing performance of Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.99$  (Problem 5). Two versions of value iteration are also appended for contrast

Algorithm (Problem 5) ( $\gamma = 0.99$ )	Runtime (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	0.37247 $\pm$ 0.0122	114617.4000 $\pm$ 187.8731	0.0%	0.0053 $\pm$ 0.0009	1517.0000 $\pm$ 162.5334	0.1398 $\pm$ 0.0067	45000.0000 $\pm$ 2280.5509	100.0%	31/31
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	0.35452 $\pm$ 0.0076	109239.9000 $\pm$ 623.2713	0.0%	0.0045 $\pm$ 0.0005	1506.9000 $\pm$ 227.1913	0.0896 $\pm$ 0.0082	30700.0000 $\pm$ 2100.0000	100.0%	31/33
Q-learning ( $\rho = 0.9, \epsilon = 0$ )	0.37097 $\pm$ 0.0538	107828.4000 $\pm$ 1384.4830	0.0%	0.0050 $\pm$ 0.0007	1494.9000 $\pm$ 151.1717	0.1082 $\pm$ 0.0769	33200.0000 $\pm$ 23990.8316	100.0%	31/100002
Q-learning ( $\rho = 0.99, \epsilon = 0$ )	0.35672 $\pm$ 0.0116	108514.1000 $\pm$ 1709.4866	0.0%	0.0056 $\pm$ 0.0012	1559.6000 $\pm$ 204.3880	0.1860 $\pm$ 0.0609	56600.0000 $\pm$ 18704.0103	100.0%	31/296
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	0.62690 $\pm$ 0.0091	147204.6000 $\pm$ 297.3829	0.0%	0.0061 $\pm$ 0.0016	1863.2000 $\pm$ 277.2716	0.2288 $\pm$ 0.0776	73857.1429 $\pm$ 25881.2043	70.0%	31/31
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	0.46012 $\pm$ 0.0137	141131.8000 $\pm$ 289.5506	0.0%	0.0055 $\pm$ 0.0011	1711.8000 $\pm$ 268.4846	0.2132 $\pm$ 0.1013	69000.0000 $\pm$ 29588.8493	40.0%	31/34
Q-learning ( $\rho = 0.9, \epsilon = 0.25$ )	0.43303 $\pm$ 0.0106	138488.2000 $\pm$ 285.2051	0.0%	0.0060 $\pm$ 0.0011	1913.9000 $\pm$ 250.9874	0.3568 $\pm$ 0.0631	114500.0000 $\pm$ 16500.0000	20.0%	31/33
Q-learning ( $\rho = 0.99, \epsilon = 0.25$ )	0.44265 $\pm$ 0.0125	137993.9000 $\pm$ 306.2292	0.0%	0.0052 $\pm$ 0.0009	1596.2000 $\pm$ 171.7066	nan $\pm$ nan	nan $\pm$ nan	0.0%	31/34
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	0.69724 $\pm$ 0.0156	214887.9000 $\pm$ 1162.4659	0.0%	0.0076 $\pm$ 0.0018	2378.4000 $\pm$ 657.2575	0.2220 $\pm$ 0.1674	105600.0000 $\pm$ 56873.8956	50.0%	31/33
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	0.65335 $\pm$ 0.0154	209485.3000 $\pm$ 618.2669	0.0%	0.0066 $\pm$ 0.0009	2187.1000 $\pm$ 314.9376	0.3104 $\pm$ 0.0783	102250.0000 $\pm$ 26309.4565	40.0%	31/32
Q-learning ( $\rho = 0.9, \epsilon = 0.5$ )	0.64415 $\pm$ 0.0115	207142.8000 $\pm$ 821.3853	0.0%	0.0075 $\pm$ 0.0024	2378.1000 $\pm$ 482.4167	nan $\pm$ nan	nan $\pm$ nan	0.0%	31/33
Q-learning ( $\rho = 0.99, \epsilon = 0.5$ )	0.65736 $\pm$ 0.0098	208470.3000 $\pm$ 881.3810	0.0%	0.0070 $\pm$ 0.0022	2325.8000 $\pm$ 714.1581	nan $\pm$ nan	nan $\pm$ nan	0.0%	31/33
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	1.29698 $\pm$ 0.0385	439024.7000 $\pm$ 2972.1627	0.0%	0.0081 $\pm$ 0.0027	2714.4000 $\pm$ 799.0636	0.6333 $\pm$ 0.2917	211000.0000 $\pm$ 97382.8967	70.0%	31/32
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	1.30140 $\pm$ 0.0282	427028.4000 $\pm$ 2753.3805	0.0%	0.0085 $\pm$ 0.0029	2850.4000 $\pm$ 849.9456	0.9404 $\pm$ 0.1961	311400.0000 $\pm$ 61856.6984	50.0%	31/33
Q-learning ( $\rho = 0.9, \epsilon = 0.75$ )	1.30231 $\pm$ 0.0223	429349.7000 $\pm$ 2556.0207	0.0%	0.0078 $\pm$ 0.0016	2563.7000 $\pm$ 461.9909	nan $\pm$ nan	nan $\pm$ nan	0.0%	31/33
Q-learning ( $\rho = 0.99, \epsilon = 0.75$ )	1.33780 $\pm$ 0.0157	441445.0000 $\pm$ 1647.8951	0.0%	0.0102 $\pm$ 0.0026	3306.3000 $\pm$ 981.7909	nan $\pm$ nan	nan $\pm$ nan	0.0%	31/32
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	3.22848 $\pm$ 0.0418	1092934.3000 $\pm$ 11159.3953	0.0%	0.0094 $\pm$ 0.0059	3136.1000 $\pm$ 1884.8837	1.9285 $\pm$ 0.6059	438500.0000 $\pm$ 204011.6418	80.0%	31/33
Q-learning ( $\rho = 0.7, \epsilon = 0.9$ )	3.27396 $\pm$ 0.0773	1101426.5000 $\pm$ 1078.4900	0.0%	0.0079 $\pm$ 0.0044	2225.1000 $\pm$ 1158.6940	1.5485 $\pm$ 0.9478	518000.0000 $\pm$ 317805.4961	60.0%	31/32
Q-learning ( $\rho = 0.9, \epsilon = 0.9$ )	3.45066 $\pm$ 0.0453	1168286.4000 $\pm$ 10996.5994	0.0%	0.0083 $\pm$ 0.0050	3944.8000 $\pm$ 1847.3530	2.5964 $\pm$ 0.1030	885000.0000 $\pm$ 170000.0000	20.0%	31/32
Q-learning ( $\rho = 0.99, \epsilon = 0.9$ )	3.70593 $\pm$ 0.0484	1253586.9000 $\pm$ 14448.1124	0.0%	0.0090 $\pm$ 0.0026	2996.3000 $\pm$ 920.5565	nan $\pm$ nan	nan $\pm$ nan	0.0%	31/34
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	13.94533 $\pm$ 0.1766	4764212.8000 $\pm$ 34616.4652	0.0%	0.0050 $\pm$ 0.0042	1618.4000 $\pm$ 1324.5059	3.0011 $\pm$ 31.967	1022900.0000 $\pm$ 1070787.3225	100.0%	31/32
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	13.99891 $\pm$ 0.2207	4795080.2000 $\pm$ 51372.6601	0.0%	0.0036 $\pm$ 0.0021	1223.7000 $\pm$ 618.9092	2.7212 $\pm$ 1.9902	932100.0000 $\pm$ 680721.5951	100.0%	31/33
Q-learning ( $\rho = 0.9, \epsilon = 1$ )	13.84819 $\pm$ 0.1945	4709311.1000 $\pm$ 47803.9988	0.0%	0.0067 $\pm$ 0.0065	2251.4000 $\pm$ 2306.8599	4.5475 $\pm$ 3.5510	1559333.3333 $\pm$ 1216611.6760	60.0%	31/32
Q-learning ( $\rho = 0.99, \epsilon = 1$ )	13.68032 $\pm$ 0.1248	4745535.3000 $\pm$ 22907.9805	0.0%	0.0060 $\pm$ 0.0057	2078.2000 $\pm$ 1974.2113	nan $\pm$ nan	nan $\pm$ nan	0.0%	31/33
Async Value Iteration	0.10508 $\pm$ 0.0038	14914.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	31/32
Value Iteration	0.13854 $\pm$ 0.0051	188718.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	31/31

**Table 40.** Comparing performance of Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.999$  (Problem 5). Two versions of value iteration are also appended for contrast

Algorithm (Problem 5) ( $\gamma = 0.999$ )	Runtime (mean $\pm$ std)	#actions (mean $\pm$ std)
---	-----------------------------	------------------------------

**Table 41.** Comparing performance of Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.99$  (Problem 7). Two versions of value iteration are also appended for contrast

Algorithm (Problem 7) ( $\gamma = 0.99$ )	Runtime (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	0.33427 $\pm$ 0.0122	105125.7000 $\pm$ 168.1048	0.0%	0.0033 $\pm$ 0.0005	1107.3000 $\pm$ 117.2610	0.0980 $\pm$ 0.0477	30100.0000 $\pm$ 1340.1219	100.0%	29/31
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	0.32748 $\pm$ 0.0138	101730.0000 $\pm$ 1430.0034	0.0%	0.0030 $\pm$ 0.0006	1015.1000 $\pm$ 129.7131	0.0700 $\pm$ 0.0045	21700.0000 $\pm$ 1100.0000	100.0%	29/31
Q-learning ( $\rho = 0.9, \epsilon = 0$ )	0.32981 $\pm$ 0.0112	102508.9000 $\pm$ 1459.2476	0.0%	0.0032 $\pm$ 0.0007	1013.0000 $\pm$ 145.9377	0.0863 $\pm$ 0.0586	27600.0000 $\pm$ 18900.7936	100.0%	29/37
Q-learning ( $\rho = 0.9, \epsilon = 0$ )	0.33552 $\pm$ 0.0117	104173.0000 $\pm$ 1482.8527	0.0%	0.0034 $\pm$ 0.0012	1029.0000 $\pm$ 118.8242	0.1318 $\pm$ 0.0789	40888.8889 $\pm$ 25431.5834	90.0%	29/35
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	0.43405 $\pm$ 0.0090	138928.2000 $\pm$ 499.2035	0.0%	0.0040 $\pm$ 0.0011	1425.5000 $\pm$ 231.0885	0.1752 $\pm$ 0.0769	57333.3333 $\pm$ 21664.1024	90.0%	29/31
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	0.41609 $\pm$ 0.0110	133855.8000 $\pm$ 361.0805	0.0%	0.0045 $\pm$ 0.0008	1314.9000 $\pm$ 212.1652	0.2415 $\pm$ 0.1086	77428.5714 $\pm$ 35362.8423	70.0%	29/30
Q-learning ( $\rho = 0.9, \epsilon = 0.25$ )	0.40879 $\pm$ 0.0130	131461.9000 $\pm$ 483.5714	0.0%	0.0036 $\pm$ 0.0011	1102.2000 $\pm$ 309.2872	nan $\pm$ nan	nan $\pm$ nan	0.0%	29/31
Q-learning ( $\rho = 0.9, \epsilon = 0.25$ )	0.40781 $\pm$ 0.0077	131072.9000 $\pm$ 596.3297	0.0%	0.0042 $\pm$ 0.0008	1219.2000 $\pm$ 187.8818	nan $\pm$ nan	nan $\pm$ nan	0.0%	29/31
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	0.64501 $\pm$ 0.0149	210120.2000 $\pm$ 559.0620	0.0%	0.0048 $\pm$ 0.0015	1592.2000 $\pm$ 523.0178	0.5552 $\pm$ 0.0608	181600.0000 $\pm$ 28302.6501	50.0%	29/31
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	0.62716 $\pm$ 0.0089	204995.1000 $\pm$ 479.7298	0.0%	0.0048 $\pm$ 0.0013	1485.7000 $\pm$ 397.7974	0.3024 $\pm$ 0.1961	100000.0000 $\pm$ 66000.0000	20.0%	29/33
Q-learning ( $\rho = 0.9, \epsilon = 0.5$ )	0.62131 $\pm$ 0.0105	202853.4000 $\pm$ 1130.9142	0.0%	0.0049 $\pm$ 0.0017	1574.0000 $\pm$ 500.7021	nan $\pm$ nan	nan $\pm$ nan	0.0%	29/31
Q-learning ( $\rho = 0.9, \epsilon = 0.5$ )	0.62282 $\pm$ 0.0129	203244.8000 $\pm$ 818.0984	0.0%	0.0044 $\pm$ 0.0012	1410.1000 $\pm$ 314.7698	nan $\pm$ nan	nan $\pm$ nan	0.0%	29/31
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	1.35667 $\pm$ 0.0148	450928.0000 $\pm$ 2101.6706	0.0%	0.0060 $\pm$ 0.0037	2345.3000 $\pm$ 1070.9297	0.5004 $\pm$ 0.2531	169428.5714 $\pm$ 85780.2127	70.0%	29/29
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	1.35292 $\pm$ 0.0245	459012.2000 $\pm$ 1877.8213	0.0%	0.0066 $\pm$ 0.0034	2197.4000 $\pm$ 1178.8474	0.6888 $\pm$ 0.2734	229750.0000 $\pm$ 94221.4811	40.0%	29/30
Q-learning ( $\rho = 0.9, \epsilon = 0.75$ )	1.34463 $\pm$ 0.0268	455792.3000 $\pm$ 1940.4911	0.0%	0.0078 $\pm$ 0.0037	2449.2000 $\pm$ 1051.0859	0.7965 $\pm$ 0.0000	266000.0000 $\pm$ 0.0000	10.0%	29/31
Q-learning ( $\rho = 0.9, \epsilon = 0.75$ )	1.35705 $\pm$ 0.0132	459658.1000 $\pm$ 1821.0783	0.0%	0.0075 $\pm$ 0.0040	2472.1000 $\pm$ 1381.1790	nan $\pm$ nan	nan $\pm$ nan	0.0%	29/29
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	4.10222 $\pm$ 0.0447	1416290.2000 $\pm$ 13406.2110	0.0%	0.0072 $\pm$ 0.0058	2373.3000 $\pm$ 1710.7167	0.6198 $\pm$ 0.5014	214400.0000 $\pm$ 37379.8607	100.0%	29/30
Q-learning ( $\rho = 0.7, \epsilon = 0.9$ )	4.09371 $\pm$ 0.0562	1403508.6000 $\pm$ 16330.4633	0.0%	0.0075 $\pm$ 0.0041	2689.1000 $\pm$ 1621.3430	1.0343 $\pm$ 0.5457	356100.0000 $\pm$ 182611.3085	100.0%	29/31
Q-learning ( $\rho = 0.9, \epsilon = 0.9$ )	4.16473 $\pm$ 0.0605	1425769.6000 $\pm$ 9162.1709	0.0%	0.0082 $\pm$ 0.0045	2627.8000 $\pm$ 1532.1414	1.6117 $\pm$ 1.1108	550250.0000 $\pm$ 378253.8004	80.0%	29/30
Q-learning ( $\rho = 0.9, \epsilon = 0.9$ )	4.22978 $\pm$ 0.0746	141110.8000 $\pm$ 14152.2690	0.0%	0.0089 $\pm$ 0.0055	2772.5000 $\pm$ 1824.4283	nan $\pm$ nan	nan $\pm$ nan	0.0%	29/30
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	17.30241 $\pm$ 0.1672	5971427.6000 $\pm$ 66264.1556	0.0%	0.0105 $\pm$ 0.0099	3899.2000 $\pm$ 4044.8257	0.8476 $\pm$ 0.3117	296100.0000 $\pm$ 105706.6223	100.0%	29/30
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	17.21702 $\pm$ 0.2234	5933093.4000 $\pm$ 54805.6557	0.0%	0.0050 $\pm$ 0.0045	1572.8000 $\pm$ 1801.3915	0.9488 $\pm$ 0.6939	325300.0000 $\pm$ 240883.4645	100.0%	29/29
Q-learning ( $\rho = 0.9, \epsilon = 1$ )	17.40180 $\pm$ 0.1782	59919735.3000 $\pm$ 40936.1980	0.0%	0.0071 $\pm$ 0.0037	2606.7000 $\pm$ 1454.0691	2.0168 $\pm$ 1.9017	698400.0000 $\pm$ 659821.2182	100.0%	29/33
Q-learning ( $\rho = 0.9, \epsilon = 1$ )	17.29415 $\pm$ 0.1359	5979988.0000 $\pm$ 4532.0784	0.0%	0.0079 $\pm$ 0.0051	2628.2000 $\pm$ 1748.4152	14.1482 $\pm$ 0.0000	4870000.0000 $\pm$ 0.0000	15.0%	29/29
Async Value Iteration	0.10354 $\pm$ 0.0032	153272.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	29/32
Value Iteration	0.15642 $\pm$ 0.0028	228344.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	29/32

**Table 42.** Comparing performance of Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.999$  (Problem 7). Two versions of value iteration are also appended for contrast

Algorithm (Problem 7) ( $\gamma = 0.999$ )	Runtime (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	0.33216 $\pm$ 0.0107	103263.5000 $\pm$ 97.8951	0.0%	0.0034 $\pm$ 0.0004	991.9000 $\pm$ 513.888	0.0905 $\pm$ 0.0071	27900.0000 $\pm$ 300.0000	100.0%	29/29
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	0.31104 $\pm$ 0.0058	98369.3000 $\pm$ 169.5149	0.0%	0.0038 $\pm$ 0.0006	1050.7000 $\pm$ 4.9000	0.0637 $\pm$ 0.0026	20100.0000 $\pm$ 300.0000	100.0%	29/31
Q-learning ( $\rho = 0.9, \epsilon = 0$ )	0.30551 $\pm$ 0.0091	96413.8000 $\pm$ 1629.3815	0.0%	0.0033 $\pm$ 0.0005	1046.6000 $\pm$ 41.6370	0.0503 $\pm$ 0.0020	16000.0000 $\pm$ 0.0000	100.0%	29/31
Q-learning ( $\rho = 0.9, \epsilon = 0$ )	0.30452 $\pm$ 0.0095	95497.2000 $\pm$ 209.7593	0.0%	0.0032 $\pm$ 0.0007	1015.1000 $\pm$ 83.5146	0.0477 $\pm$ 0.0031	14700.0000 $\pm$ 458.2576	100.0%	29/30
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	0.42343 $\pm$ 0.0088	136151.9000 $\pm$ 548.4530	0.0%	0.0047 $\pm$ 0.0008	1411.1000 $\pm$ 257.0599	0.0966 $\pm$ 0.0037	30800.0000 $\pm$ 400.0000	100.0%	29/31
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	0.40466 $\pm$ 0.0088	130875.1000 $\pm$ 687.4162	0.0%	0.0046 $\pm$ 0.0029	1373.0000 $\pm$ 397.5777	0.0715 $\pm$ 0.0023	22600.0000 $\pm$ 889.8979	100.0%	29/29
Q-learning ( $\rho = 0.9, \epsilon = 0.25$ )	0.40134 $\pm$ 0.0112	128105.8000 $\pm$ 324.8128	0.0%	0.0039 $\pm$ 0.0007	1251.4000 $\pm$ 130.5214	0.0579 $\pm$ 0.0032	18100.0000 $\pm$ 300.0000	100.0%	29/29
Q-learning ( $\rho = 0.9, \epsilon = 0.25$ )	0.39706 $\pm$ 0.0101	127570.9000 $\pm$ 316.5042	0.0%	0.0039 $\pm$ 0.0009	1311.8000 $\pm$ 203.7094	0.0518 $\pm$ 0.0015	16800.0000 $\pm$ 400.0000	100.0%	29/29
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	0.63257 $\pm$ 0.0182	205580.9000 $\pm$ 1404.5667	0.0%	0.0052 $\pm$ 0.0015	1622.2000 $\pm$ 578.2848	0.1113 $\pm$ 0.0068	35800.0000 $\pm$ 600.0000	100.0%	29/29
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	0.62011 $\pm$ 0.0124	200313.1000 $\pm$ 1089.6123	0.0%	0.0054 $\pm$ 0.0015	1697.9000 $\pm$ 377.1388	0.1848 $\pm$ 0.0660	27200.0000 $\pm$ 300.0000	100.0%	29/30
Q-learning ( $\rho = 0.9, \epsilon = 0.5$ )	0.60459 $\pm$ 0.0116	199227.1000 $\pm$ 805.0397	0.0%	0.0041 $\pm$ 0.0013	1376.9000 $\pm$ 330.9065	0.0686 $\pm$ 0.0038	22500.0000 $\pm$ 500.0000	100.0%	29/29
Q-learning ( $\rho = 0.9, \epsilon = 0.5$ )	0.60547 $\pm$ 0.0120	199838.4000 $\pm$ 1353.3196	0.0%	0.0054 $\pm$ 0.0017	1721.2000 $\pm$ 528.9037	0.0623 $\pm$ 0.0038	20200.0000 $\pm$ 600.0000	100.0%	29/29
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	1.35722 $\pm$ 0.0182	455451.6000 $\pm$ 3052.3085	0.0%	0.0063 $\pm$ 0.0040	2006.3000 $\pm$ 1203.4328	0.1411 $\pm$ 0.0052	47300.0000 $\pm$ 900.0000	100.0%	29/29
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	1.33919 $\pm$ 0.0179	451601.6000 $\pm$ 3530.7105	0.0%	0.0060 $\pm$ 0.0035	2007.4000 $\pm$ 1127.2885	0.1116 $\pm$ 0.0055	37000.0000 $\pm$ 1095.4451	100.0%	29/29
Q-learning ( $\rho = 0.9, \epsilon = 0.75$ )	1.37123 $\pm$ 0.0189	460888.6000 $\pm$ 2646.9526	0.0%	0.0073 $\pm$ 0.0028	2533.5000 $\pm$ 1178.8222	0.0933 $\pm$ 0.0061	31900.0000 $\pm$ 1135.7817	100.0%	29/30
Q-learning ( $\rho = 0.9, \epsilon = 0.75$ )	1.37888 $\pm$ 0.0150	467227.8000 $\pm$ 3335.9965	0.0%	0.0052 $\pm$ 0.0020	1786.7000 $\pm$ 533.3095	0.0851 $\pm$ 0.0035	28400.0000 $\pm$ 663.3250	100.0%	29/30
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	4.15838 $\pm$ 0.0088	1426565.1000 $\pm$ 18949.2330	10.0%	0.0072 $\pm$ 0.0041	2383.6000 $\pm$ 1194.7452	0.2110 $\pm$ 0.0190	70900.0000 $\pm$ 5374.9419	100.0%	29/29
Q-learning ( $\rho = 0.7, \epsilon = 0.9$ )	3.40145 $\pm$ 0.2290	1168688.8000 $\pm$ 23906.4924	80.0%	0.0053 $\pm$ 0.0045	1952.8000 $\pm$ 1346.8030	0.1548 $\pm$ 0.0165	32900.0000 $\pm$ 3986.2293	100.0%	29/29
Q-learning ( $\rho = 0.9, \epsilon = 0.9$ )	2.64154 $\pm$ 0.4819	907939.3000 $\pm$ 167557.7797	100.0%	0.0071 $\pm$ 0.0051	2274.9000 $\pm$ 1836.3152	0.1330 $\pm$ 0.0121	45500.0000 $\pm$ 2692.5824	100.0%	29/29
Q-learning ( $\rho = 0.9, \epsilon = 0.9$ )	1.76411 $\pm$ 0.4392	606538.8000 $\pm$ 153380.7036	100.0%	0.0048 $\pm$ 0.0020	1577.6000 $\pm$ 557.6015	0.1184 $\pm$ 0.0113	41700.0000 $\pm$ 2609.5977	100.0%	29/30
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	3.02736 $\pm$ 0.3839	1041897.5000 $\pm$ 133034.0531	100.0%	0.0050 $\pm$ 0.0037	1883.4000 $\pm$ 1331.9621	0.0406 $\pm$ 0.0763	207400.0000 $\pm$ 24670.6303	100.0%	29/31
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	2.60879 $\pm$ 0.5753	901490.4000 $\pm$ 202265.8853	100.0%	0.0082 $\pm$ 0.0077	2664.7000 $\pm$ 2314.6502	0.4190 $\pm$ 0.0814	144900.0000 $\pm$ 27005.3698	100.0%	29/29
Q-learning ( $\rho = 0.9, \epsilon = 1$ )	1.98026 $\pm$ 0.3646	680915.2000 $\pm$ 125997.0656	100.0%	0.0108 $\pm$ 0.0084	3578.7000 $\pm$ 2836.1716	0.3622 $\pm$ 0.0473	122700.0000 $\pm$ 14036.0251	100.0%	29/29
Q-learning ( $\rho = 0.9, \epsilon = 1$ )	2.26373 $\pm$ 0.4705	757744.1000 $\pm$ 151013.5598	100.0%	0.0101 $\pm$ 0.0096	3579.8000 $\pm$ 3820.0271	0.3187 $\pm$ 0.0480	109000.0000 $\pm$ 16619.2659	100.0%	29/29
Async Value Iteration	0.08177 $\pm$ 0.0028	125120.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	29/29
Value Iteration	0.12583 $\pm$ 0.0071	181424.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	29/29

**Table 43.** Comparing performance of Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.99$  (Problem 8). Two versions of value iteration are also appended for contrast

Algorithm (Problem 8) ( $\gamma = 0.99$ )	Runtime (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal 
--	-----------------------------	------------------------------	-------------	-------------------

**Table 44.** Comparing performance of Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.999$  (Problem 8). Two versions of value iteration are also appended for contrast

Algorithm (Problem 8) ( $\gamma = 0.999$ )	Runtime (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	0.01766 $\pm$ 0.0023	15215.5000 $\pm$ 10.1415	0.0%	0.0010 $\pm$ 0.0000	39.4000 $\pm$ 1.2000	0.0032 $\pm$ 0.0004	1000.0000 $\pm$ 0.0000	100.0%	6/6
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	0.01846 $\pm$ 0.0017	15174.0000 $\pm$ 4.1473	0.0%	0.0000 $\pm$ 0.0000	39.6000 $\pm$ 1.8000	0.0034 $\pm$ 0.0007	1000.0000 $\pm$ 0.0000	100.0%	6/6
Q-learning ( $\rho = 0.9, \epsilon = 0$ )	0.05209 $\pm$ 0.0035	15185.0000 $\pm$ 78.5519	0.0%	0.0010 $\pm$ 0.0000	39.0000 $\pm$ 0.0000	0.0035 $\pm$ 0.0005	1000.0000 $\pm$ 0.0000	100.0%	6/6
Q-learning ( $\rho = 0.99, \epsilon = 0$ )	0.01063 $\pm$ 0.0052	3200.9000 $\pm$ 1536.6980	100.0%	0.0010 $\pm$ 0.0000	39.5000 $\pm$ 1.0247	0.0039 $\pm$ 0.0015	1200.0000 $\pm$ 600.0000	100.0%	6/7
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	0.01759 $\pm$ 0.0087	5703.0000 $\pm$ 2865.5200	100.0%	0.0010 $\pm$ 0.0000	50.5000 $\pm$ 9.7082	0.0033 $\pm$ 0.0007	1000.0000 $\pm$ 0.0000	100.0%	6/6
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	0.01216 $\pm$ 0.0064	4103.7000 $\pm$ 2342.9184	100.0%	0.0000 $\pm$ 0.0000	47.5000 $\pm$ 8.0654	0.0030 $\pm$ 0.0006	1000.0000 $\pm$ 0.0000	100.0%	6/6
Q-learning ( $\rho = 0.9, \epsilon = 0.25$ )	0.00946 $\pm$ 0.0065	3092.1000 $\pm$ 3194.1199	100.0%	0.0012 $\pm$ 0.0002	47.4000 $\pm$ 8.2849	0.0035 $\pm$ 0.0006	1000.0000 $\pm$ 0.0000	100.0%	6/6
Q-learning ( $\rho = 0.99, \epsilon = 0.25$ )	0.00529 $\pm$ 0.0035	1602.9000 $\pm$ 1018.7575	100.0%	0.0010 $\pm$ 0.0000	48.0000 $\pm$ 7.8102	0.0034 $\pm$ 0.0005	1000.0000 $\pm$ 0.0000	100.0%	6/6
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	0.00905 $\pm$ 0.0076	2905.5000 $\pm$ 2425.9665	100.0%	0.0010 $\pm$ 0.0000	52.8000 $\pm$ 20.3263	0.0035 $\pm$ 0.0012	1100.0000 $\pm$ 300.0000	100.0%	6/6
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	0.00372 $\pm$ 0.0010	1306.9000 $\pm$ 456.8458	100.0%	0.0010 $\pm$ 0.0000	54.4000 $\pm$ 15.8569	0.0034 $\pm$ 0.0007	1000.0000 $\pm$ 0.0000	100.0%	6/6
Q-learning ( $\rho = 0.9, \epsilon = 0.5$ )	0.00515 $\pm$ 0.0054	1506.5000 $\pm$ 1206.1261	100.0%	0.0010 $\pm$ 0.0000	46.8000 $\pm$ 11.8474	0.0030 $\pm$ 0.0005	1000.0000 $\pm$ 0.0000	100.0%	6/6
Q-learning ( $\rho = 0.99, \epsilon = 0.5$ )	0.00411 $\pm$ 0.0030	1406.5000 $\pm$ 915.7572	100.0%	0.0000 $\pm$ 0.0000	56.2000 $\pm$ 10.8517	0.0031 $\pm$ 0.0003	1000.0000 $\pm$ 0.0000	100.0%	6/6
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	0.00819 $\pm$ 0.0027	1912.7000 $\pm$ 702.8027	100.0%	0.0000 $\pm$ 0.0000	65.1000 $\pm$ 26.7000	0.0033 $\pm$ 0.0007	1000.0000 $\pm$ 0.0000	100.0%	6/6
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	0.00425 $\pm$ 0.0016	1410.6000 $\pm$ 487.8080	100.0%	0.0010 $\pm$ 0.0000	63.2000 $\pm$ 31.5144	0.0032 $\pm$ 0.0004	1000.0000 $\pm$ 0.0000	100.0%	6/6
Q-learning ( $\rho = 0.9, \epsilon = 0.75$ )	0.00386 $\pm$ 0.0021	1218.3000 $\pm$ 596.5878	100.0%	0.0000 $\pm$ 0.0000	50.4000 $\pm$ 21.1858	0.0032 $\pm$ 0.0011	1000.0000 $\pm$ 0.0000	100.0%	6/6
Q-learning ( $\rho = 0.99, \epsilon = 0.75$ )	0.00419 $\pm$ 0.0019	1309.7000 $\pm$ 455.9820	100.0%	0.0010 $\pm$ 0.0000	44.0000 $\pm$ 26.6871	0.0033 $\pm$ 0.0011	1100.0000 $\pm$ 300.0000	100.0%	6/6
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	0.00768 $\pm$ 0.0027	2624.7000 $\pm$ 918.9460	100.0%	0.0010 $\pm$ 0.0000	58.6000 $\pm$ 25.8116	0.0029 $\pm$ 0.0003	1000.0000 $\pm$ 0.0000	100.0%	6/7
Q-learning ( $\rho = 0.7, \epsilon = 0.9$ )	0.00620 $\pm$ 0.0027	2027.5000 $\pm$ 884.4308	100.0%	0.0010 $\pm$ 0.0000	48.2000 $\pm$ 27.4401	0.0031 $\pm$ 0.0005	1000.0000 $\pm$ 0.0000	100.0%	6/6
Q-learning ( $\rho = 0.9, \epsilon = 0.9$ )	0.00440 $\pm$ 0.0023	1416.8000 $\pm$ 662.1282	100.0%	0.0000 $\pm$ 0.0000	55.3000 $\pm$ 31.4517	0.0031 $\pm$ 0.0007	1000.0000 $\pm$ 0.0000	100.0%	6/8
Q-learning ( $\rho = 0.99, \epsilon = 0.9$ )	0.00406 $\pm$ 0.0023	1336.2000 $\pm$ 471.4076	100.0%	0.0010 $\pm$ 0.0000	38.8000 $\pm$ 18.9832	0.0029 $\pm$ 0.0006	1000.0000 $\pm$ 0.0000	100.0%	6/6
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	0.01071 $\pm$ 0.0033	3640.6000 $\pm$ 914.5519	100.0%	0.0010 $\pm$ 0.0000	24.0000 $\pm$ 12.2393	0.0033 $\pm$ 0.0017	1100.0000 $\pm$ 300.0000	100.0%	6/6
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	0.00691 $\pm$ 0.0027	2357.1000 $\pm$ 906.2989	100.0%	0.0010 $\pm$ 0.0000	46.4000 $\pm$ 30.9587	0.0031 $\pm$ 0.0003	1000.0000 $\pm$ 0.0000	100.0%	6/6
Q-learning ( $\rho = 0.9, \epsilon = 1$ )	0.00633 $\pm$ 0.0026	2130.0000 $\pm$ 832.7556	100.0%	0.0010 $\pm$ 0.0000	43.8000 $\pm$ 42.2488	0.0033 $\pm$ 0.0005	1100.0000 $\pm$ 300.0000	100.0%	6/6
Q-learning ( $\rho = 0.99, \epsilon = 1$ )	0.00321 $\pm$ 0.0011	1123.8000 $\pm$ 397.2692	100.0%	0.0010 $\pm$ 0.0000	39.0000 $\pm$ 27.1735	0.0028 $\pm$ 0.0006	1000.0000 $\pm$ 0.0000	100.0%	6/6
Async Value Iteration	0.00325 $\pm$ 0.0004	1224.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	6/6
Value Iteration	0.00489 $\pm$ 0.0005	1836.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	6/6

**Table 45.** Comparing performance of Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.99$  (Problem 9). Two versions of value iteration are also appended for contrast

Algorithm (Problem 9) ( $\gamma = 0.99$ )	Runtime (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	0.83249 $\pm$ 0.0195	264739.0000 $\pm$ 263.9455	0.0%	0.0048 $\pm$ 0.0013	1665.4000 $\pm$ 512.5989	2.2556 $\pm$ 0.0078	71100.0000 $\pm$ 2947.8806	100.0%	73/77
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	0.83514 $\pm$ 0.0599	255189.2000 $\pm$ 2224.9005	0.0%	0.0047 $\pm$ 0.0024	1359.1000 $\pm$ 475.6046	0.1622 $\pm$ 0.0120	9400.0000 $\pm$ 2015.3394	100.0%	73/10002
Q-learning ( $\rho = 0.9, \epsilon = 0$ )	0.82218 $\pm$ 0.0200	261503.5000 $\pm$ 3559.2702	0.0%	0.0047 $\pm$ 0.0021	1303.8000 $\pm$ 598.0062	0.2267 $\pm$ 0.1634	72666.6667 $\pm$ 53441.5569	90.0%	75/210
Q-learning ( $\rho = 0.99, \epsilon = 0$ )	0.84797 $\pm$ 0.0197	268361.8000 $\pm$ 2379.4391	0.0%	0.0040 $\pm$ 0.0012	1213.7000 $\pm$ 379.1628	0.2129 $\pm$ 0.0995	67000.0000 $\pm$ 32000.0000	20.0%	78/333
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	1.09374 $\pm$ 0.0134	354101.8000 $\pm$ 595.4437	0.0%	0.0061 $\pm$ 0.0028	1890.3000 $\pm$ 947.1300	3.0363 $\pm$ 0.0865	98900.0000 $\pm$ 26032.4797	100.0%	73/77
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	1.06019 $\pm$ 0.0180	343039.1000 $\pm$ 667.9831	0.0%	0.0048 $\pm$ 0.0041	1465.0000 $\pm$ 1181.8844	0.3598 $\pm$ 0.1795	11720.0000 $\pm$ 57875.8834	100.0%	73/77
Q-learning ( $\rho = 0.9, \epsilon = 0.25$ )	1.05142 $\pm$ 0.0180	340876.0000 $\pm$ 678.3694	0.0%	0.0034 $\pm$ 0.0015	1193.4000 $\pm$ 514.9717	0.3378 $\pm$ 0.1401	109000.0000 $\pm$ 46566.0821	100.0%	73/2626
Q-learning ( $\rho = 0.99, \epsilon = 0.25$ )	1.04891 $\pm$ 0.0169	342443.7000 $\pm$ 1504.2654	0.0%	0.0100 $\pm$ 0.0055	3310.6000 $\pm$ 1792.9804	0.3729 $\pm$ 0.2390	121200.0000 $\pm$ 75744.0427	100.0%	73/482
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	1.67280 $\pm$ 0.0191	542591.0000 $\pm$ 717.5992	0.0%	0.0075 $\pm$ 0.0050	3367.0000 $\pm$ 1941.9683	0.5089 $\pm$ 0.2927	91700.0000 $\pm$ 113200.8052	100.0%	73/77
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	1.62994 $\pm$ 0.0320	532254.6000 $\pm$ 1110.4212	0.0%	0.0091 $\pm$ 0.0044	3013.0000 $\pm$ 1418.1769	0.3953 $\pm$ 0.1658	131400.0000 $\pm$ 52442.7307	100.0%	73/75
Q-learning ( $\rho = 0.9, \epsilon = 0.5$ )	1.62375 $\pm$ 0.0113	532702.9000 $\pm$ 1277.0390	0.0%	0.0104 $\pm$ 0.0071	3473.3000 $\pm$ 2392.1237	0.5561 $\pm$ 0.2936	180500.0000 $\pm$ 93570.5616	100.0%	73/1215
Q-learning ( $\rho = 0.99, \epsilon = 0.5$ )	1.70275 $\pm$ 0.1623	532676.4000 $\pm$ 1119.3726	0.0%	0.0081 $\pm$ 0.0037	2570.8000 $\pm$ 1323.9925	0.4862 $\pm$ 0.3775	147000.0000 $\pm$ 113822.6691	100.0%	73/1190
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	3.17534 $\pm$ 0.0379	1170012.0000 $\pm$ 5286.8401	0.0%	0.0202 $\pm$ 0.0082	6715.7000 $\pm$ 2629.9858	0.6141 $\pm$ 0.2729	202000.0000 $\pm$ 92884.8863	100.0%	73/78
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	3.46982 $\pm$ 0.0314	1166264.4000 $\pm$ 3635.3480	0.0%	0.0166 $\pm$ 0.0090	5640.8000 $\pm$ 2844.1041	0.4409 $\pm$ 0.1569	149900.0000 $\pm$ 54725.5882	100.0%	73/75
Q-learning ( $\rho = 0.9, \epsilon = 0.75$ )	3.53218 $\pm$ 0.1073	1171866.5000 $\pm$ 2795.7716	0.0%	0.0154 $\pm$ 0.0091	4989.2000 $\pm$ 2665.4561	0.4960 $\pm$ 0.3073	166900.0000 $\pm$ 103700.0000	100.0%	73/134
Q-learning ( $\rho = 0.99, \epsilon = 0.75$ )	3.52646 $\pm$ 0.0584	1174115.7000 $\pm$ 4024.4103	0.0%	0.0144 $\pm$ 0.0078	5075.6000 $\pm$ 2696.6880	0.9399 $\pm$ 0.8252	310200.0000 $\pm$ 262035.8909	100.0%	73/396
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	9.95711 $\pm$ 0.1049	338343.0000 $\pm$ 13848.9996	0.0%	0.0370 $\pm$ 0.0120	12458.1000 $\pm$ 5189.9189	0.9297	91700.0000 $\pm$ 113200.8052	100.0%	73/75
Q-learning ( $\rho = 0.7, \epsilon = 0.9$ )	10.06367 $\pm$ 0.0805	3405791.6000 $\pm$ 19143.3199	0.0%	0.0350 $\pm$ 0.0145	12192.7000 $\pm$ 4719.8426	0.8740 $\pm$ 0.3548	296700.0000 $\pm$ 122547.9906	100.0%	73/1184
Q-learning ( $\rho = 0.9, \epsilon = 0.9$ )	10.12019 $\pm$ 0.1112	3429444.8000 $\pm$ 17535.3803	0.0%	0.0365 $\pm$ 0.0253	12366.9000 $\pm$ 8340.2219	0.9264 $\pm$ 0.3555	311400.0000 $\pm$ 118766.3252	100.0%	73/3031
Q-learning ( $\rho = 0.99, \epsilon = 0.9$ )	10.07967 $\pm$ 0.0874	3432788.3000 $\pm$ 12238.1818	0.0%	0.0306 $\pm$ 0.0195	10482.2000 $\pm$ 6587.8251	0.8221 $\pm$ 0.4847	281200.0000 $\pm$ 167865.3031	100.0%	73/75
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	25.73975 $\pm$ 0.1861	8858962.2000 $\pm$ 16766.6485	0.0%	0.1131 $\pm$ 0.1542	88799.6000 $\pm$ 5374.2204	2.2514 $\pm$ 0.6671	770100.0000 $\pm$ 227671.9657	100.0%	73/74
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	25.85669 $\pm$ 0.2058	8866683.8000 $\pm$ 10524.3605	0.0%	0.1203 $\pm$ 0.1520	41029.9000 $\pm$ 50871.7900	2.0149 $\pm$ 0.5291	697000.0000 $\pm$ 181209.8231	100.0%	73/80
Q-learning ( $\rho = 0.9, \epsilon = 1$ )	25.93604 $\pm$ 0.2182	8866027.2000 $\pm$ 10242.8182	0.0%	0.1469 $\pm$ 0.0673	51328.8000 $\pm$ 24185.7968	1.8719 $\pm$ 0.6408	642100.0000 $\pm$ 216747.5259	100.0%	73/78
Q-learning ( $\rho = 0.99, \epsilon = 1$ )	28.16589 $\pm$ 1.7915	8866733.0000 $\pm$ 15794.0405	0.0%	0.1938 $\pm$ 0.1284	62289.1000 $\pm$ 42245.4083	1.4471 $\pm$ 0.2528	468700.0000 $\pm$ 95596.0773	100.0%	73/75
Async Value Iteration	0.08866 $\pm$ 0.0045	85158.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	73/77
Value Iteration	0.16040 $\pm$ 0.0039	180774.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	73/75

**Table 46.** Comparing performance of Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.999$  (Problem 9). Two versions of value iteration are also appended for contrast

Algorithm (Problem 9) ( $\gamma = 0.999$ )	Runtime (mean $\pm$ std)	#actions (mean $\pm$
---	-----------------------------	-------------------------

**Table 47.** Comparing performance of Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.99$  (Problem 12). Two versions of value iteration are also appended for contrast

Algorithm (Problem 12) ( $\gamma = 0.99$ )	Runtime (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	1.35049 $\pm$ 0.1206	349487.5000 $\pm$ 371.1585	0.0%	0.0154 $\pm$ 0.0096	3049.8000 $\pm$ 975.2360	0.4556 $\pm$ 0.0755	116300.0000 $\pm$ 2282.5424	100.0%	91/220
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	1.21918 $\pm$ 0.0737	330855.7000 $\pm$ 1200.4179	0.0%	0.0093 $\pm$ 0.0027	2602.5000 $\pm$ 715.6272	0.3117 $\pm$ 0.0414	84100.0000 $\pm$ 3112.8765	100.0%	91/99
Q-learning ( $\rho = 0.9, \epsilon = 0$ )	1.18828 $\pm$ 0.0713	330480.3000 $\pm$ 2305.1930	0.0%	0.0092 $\pm$ 0.0034	2725.0000 $\pm$ 1024.8644	0.3035 $\pm$ 0.1568	86700.0000 $\pm$ 40456.5388	100.0%	93/99
Q-learning ( $\rho = 0.99, \epsilon = 0$ )	1.18778 $\pm$ 0.0537	334367.3000 $\pm$ 1915.6908	0.0%	0.0081 $\pm$ 0.0027	2486.8000 $\pm$ 815.5651	0.2972 $\pm$ 0.2505	85900.0000 $\pm$ 72693.1221	100.0%	96/100002
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	1.57223 $\pm$ 0.0228	452901.9000 $\pm$ 558.4995	0.0%	0.0116 $\pm$ 0.0089	3039.6000 $\pm$ 2594.0057	0.5420 $\pm$ 0.0761	154800.0000 $\pm$ 21398.1131	100.0%	91/614
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	1.49643 $\pm$ 0.0211	422825.5000 $\pm$ 286.9872	0.0%	0.0099 $\pm$ 0.0043	3055.3000 $\pm$ 1180.2092	0.6327 $\pm$ 0.2000	153000.0000 $\pm$ 23615.4342	100.0%	91/94
Q-learning ( $\rho = 0.9, \epsilon = 0.25$ )	1.58308 $\pm$ 0.1846	426325.3000 $\pm$ 997.9812	0.0%	0.0116 $\pm$ 0.0070	2836.6000 $\pm$ 1404.7186	0.5289 $\pm$ 0.2234	136100.0000 $\pm$ 49902.8055	100.0%	91/95
Q-learning ( $\rho = 0.99, \epsilon = 0.25$ )	1.47583 $\pm$ 0.0345	427504.5000 $\pm$ 1672.9297	0.0%	0.0092 $\pm$ 0.0057	2317.6000 $\pm$ 973.4065	0.6362 $\pm$ 0.3604	184333.3333 $\pm$ 105651.4185	60.0%	91/143
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	2.26246 $\pm$ 0.0669	663258.0000 $\pm$ 1233.7467	0.0%	0.0173 $\pm$ 0.0077	5240.3000 $\pm$ 2291.8347	0.6606 $\pm$ 0.1448	196500.0000 $\pm$ 38933.9184	100.0%	91/95
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	2.17321 $\pm$ 0.0559	64657.7000 $\pm$ 1235.6505	0.0%	0.0156 $\pm$ 0.0062	4726.7000 $\pm$ 1790.8730	0.7594 $\pm$ 0.4649	228900.0000 $\pm$ 142064.6332	100.0%	91/96
Q-learning ( $\rho = 0.9, \epsilon = 0.5$ )	2.20964 $\pm$ 0.0344	642264.3000 $\pm$ 1757.2241	0.0%	0.0134 $\pm$ 0.0083	3833.0000 $\pm$ 2582.2064	1.0991 $\pm$ 0.7577	317857.1429 $\pm$ 215001.2814	70.0%	91/93
Q-learning ( $\rho = 0.99, \epsilon = 0.5$ )	2.19196 $\pm$ 0.0347	640181.9000 $\pm$ 2028.4827	0.0%	0.0168 $\pm$ 0.0107	4567.4000 $\pm$ 2815.3857	1.5732 $\pm$ 0.0000	468000.0000 $\pm$ 0.0000	10.0%	91/541
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	4.43949 $\pm$ 0.1305	1345399.1000 $\pm$ 4435.1527	0.0%	0.0272 $\pm$ 0.0142	7816.0000 $\pm$ 3092.8578	0.7407 $\pm$ 0.1363	220800.0000 $\pm$ 30514.9144	100.0%	91/94
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	4.25523 $\pm$ 0.0747	1336149.6000 $\pm$ 4826.8964	0.0%	0.0279 $\pm$ 0.0157	8969.9000 $\pm$ 5389.6546	0.7795 $\pm$ 0.3274	246800.0000 $\pm$ 99628.1085	100.0%	91/94
Q-learning ( $\rho = 0.9, \epsilon = 0.75$ )	4.24001 $\pm$ 0.0731	1334894.4000 $\pm$ 4073.3894	0.0%	0.0310 $\pm$ 0.0157	10006.7000 $\pm$ 4786.0015	1.5722 $\pm$ 0.9591	491800.0000 $\pm$ 297310.2084	100.0%	91/96
Q-learning ( $\rho = 0.99, \epsilon = 0.75$ )	4.28582 $\pm$ 0.0528	1335586.1000 $\pm$ 4076.2827	0.0%	0.0261 $\pm$ 0.0060	8509.5000 $\pm$ 1956.4541	2.4463 $\pm$ 1.4139	763000.0000 $\pm$ 424299.8537	30.0%	91/94
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	10.90180 $\pm$ 0.5767	370259.7000 $\pm$ 18502.6419	0.0%	0.0586 $\pm$ 0.0249	20067.2000 $\pm$ 8075.6777	0.9018 $\pm$ 0.1545	367800.0000 $\pm$ 46801.6917	100.0%	91/95
Q-learning ( $\rho = 0.7, \epsilon = 0.9$ )	10.73192 $\pm$ 0.0745	3719210.1000 $\pm$ 25247.8884	0.0%	0.0323 $\pm$ 0.0182	11336.6000 $\pm$ 6072.4067	1.0070 $\pm$ 0.3160	353100.0000 $\pm$ 109684.5021	100.0%	91/95
Q-learning ( $\rho = 0.9, \epsilon = 0.9$ )	10.76368 $\pm$ 0.1441	3716656.7000 $\pm$ 17669.6110	0.0%	0.0499 $\pm$ 0.0196	17346.3000 $\pm$ 6210.4451	2.4042 $\pm$ 2.7515	833100.0000 $\pm$ 949126.5932	100.0%	91/644
Q-learning ( $\rho = 0.99, \epsilon = 0.9$ )	10.77006 $\pm$ 0.1187	3737648.0000 $\pm$ 22348.5651	0.0%	0.0415 $\pm$ 0.0264	14317.1000 $\pm$ 8408.5730	4.0375 $\pm$ 2.2799	149677.7778 $\pm$ 798277.6985	90.0%	91/508
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	25.58747 $\pm$ 0.4112	894929.4000 $\pm$ 8155.9487	0.0%	0.2668 $\pm$ 0.2044	91727.6000 $\pm$ 74593.1546	5.7067 $\pm$ 1.9759	2923700.0000 $\pm$ 669699.1043	100.0%	91/95
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	25.40030 $\pm$ 0.2116	8952399.5000 $\pm$ 7496.6479	0.0%	0.3933 $\pm$ 0.4682	139007.4000 $\pm$ 164452.4481	4.8250 $\pm$ 0.9852	1681300.0000 $\pm$ 339919.4169	100.0%	91/94
Q-learning ( $\rho = 0.9, \epsilon = 1$ )	25.38862 $\pm$ 0.2153	8947227.8000 $\pm$ 6037.7618	0.0%	0.4111 $\pm$ 0.2980	146740.3000 $\pm$ 105196.9891	4.3187 $\pm$ 1.0141	1525900.0000 $\pm$ 358173.8265	100.0%	91/867
Q-learning ( $\rho = 0.99, \epsilon = 1$ )	25.44586 $\pm$ 0.1526	8949491.3000 $\pm$ 7095.9280	0.0%	0.3717 $\pm$ 0.4225	131063.6000 $\pm$ 17018.7102	6.5359 $\pm$ 1.3032	2901600.0000 $\pm$ 441303.8593	100.0%	91/94
Async Value Iteration	0.13077 $\pm$ 0.0088	15850.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	91/96
Value Iteration	0.22881 $\pm$ 0.0077	27693.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	91/94

**Table 48.** Comparing performance of Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.999$  (Problem 12). Two versions of value iteration are also appended for contrast

Algorithm (Problem 12) ( $\gamma = 0.999$ )	Runtime (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	1.21094 $\pm$ 0.0363	34406.1000 $\pm$ 102.2814	0.0%	0.0084 $\pm$ 0.0012	2431.0000 $\pm$ 110.7729	0.4106 $\pm$ 0.0192	117600.0000 $\pm$ 1685.2300	100.0%	91/91
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	1.10425 $\pm$ 0.0174	329271.1000 $\pm$ 485.7095	0.0%	0.0112 $\pm$ 0.0019	3345.3000 $\pm$ 552.7379	0.2849 $\pm$ 0.0118	83300.0000 $\pm$ 2902.4984	100.0%	91/94
Q-learning ( $\rho = 0.9, \epsilon = 0$ )	1.10495 $\pm$ 0.0342	314963.3000 $\pm$ 2060.9715	0.0%	0.0098 $\pm$ 0.0019	2861.3000 $\pm$ 469.4173	0.2285 $\pm$ 0.0140	64300.0000 $\pm$ 1100.0000	100.0%	91/3222
Q-learning ( $\rho = 0.99, \epsilon = 0$ )	1.13159 $\pm$ 0.0396	312648.0000 $\pm$ 1307.9674	0.0%	0.0112 $\pm$ 0.0040	2965.7000 $\pm$ 362.5336	0.2073 $\pm$ 0.0200	57800.0000 $\pm$ 871.7798	100.0%	91/11470
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	1.55630 $\pm$ 0.0174	44708.9000 $\pm$ 672.7366	0.0%	0.0131 $\pm$ 0.0072	3785.6000 $\pm$ 2041.8368	0.4170 $\pm$ 0.0217	128900.0000 $\pm$ 2022.3748	100.0%	91/91
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	1.47309 $\pm$ 0.0541	42557.3000 $\pm$ 477.7305	0.0%	0.0074 $\pm$ 0.0016	2284.2000 $\pm$ 457.0936	0.3191 $\pm$ 0.0183	93600.0000 $\pm$ 1356.4660	100.0%	91/91
Q-learning ( $\rho = 0.9, \epsilon = 0.25$ )	1.45958 $\pm$ 0.0953	415118.0000 $\pm$ 1083.8050	0.0%	0.0115 $\pm$ 0.0070	3451.3000 $\pm$ 2222.0212	0.2585 $\pm$ 0.0230	74000.0000 $\pm$ 894.4272	100.0%	91/93
Q-learning ( $\rho = 0.99, \epsilon = 0.25$ )	1.45451 $\pm$ 0.0305	412530.3000 $\pm$ 690.9857	0.0%	0.0119 $\pm$ 0.0070	3408.0000 $\pm$ 1994.6657	0.2923 $\pm$ 0.0160	67400.0000 $\pm$ 4356.4660	100.0%	91/91
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	2.45474 $\pm$ 0.0475	654907.4000 $\pm$ 1405.2715	0.0%	0.0180 $\pm$ 0.0106	4985.3000 $\pm$ 2842.1437	0.4880 $\pm$ 0.0205	145700.0000 $\pm$ 1900.0000	100.0%	91/91
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	2.17102 $\pm$ 0.0289	640450.2000 $\pm$ 1371.2788	0.0%	0.0158 $\pm$ 0.0057	4673.3000 $\pm$ 1744.4904	0.3724 $\pm$ 0.0069	111100.0000 $\pm$ 1744.7727	100.0%	91/92
Q-learning ( $\rho = 0.9, \epsilon = 0.5$ )	2.16785 $\pm$ 0.0543	631816.6000 $\pm$ 1227.9894	0.0%	0.0175 $\pm$ 0.0071	5026.0000 $\pm$ 1940.3128	0.3151 $\pm$ 0.0172	90500.0000 $\pm$ 1204.1595	100.0%	91/91
Q-learning ( $\rho = 0.99, \epsilon = 0.5$ )	2.17225 $\pm$ 0.0409	62577.9000 $\pm$ 1655.9166	0.0%	0.0151 $\pm$ 0.0069	4572.4000 $\pm$ 1658.9468	0.2787 $\pm$ 0.0118	82900.0000 $\pm$ 1300.0000	100.0%	91/93
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	4.47757 $\pm$ 0.1016	134527.5000 $\pm$ 3977.2152	0.0%	0.0328 $\pm$ 0.0141	9890.0000 $\pm$ 4365.4134	0.6339 $\pm$ 0.0323	188600.0000 $\pm$ 2120.2510	100.0%	91/91
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	4.43670 $\pm$ 0.1067	1333923.2000 $\pm$ 4511.8552	0.0%	0.0229 $\pm$ 0.0132	7162.9000 $\pm$ 4239.4070	0.4808 $\pm$ 0.0169	146600.0000 $\pm$ 3104.8349	100.0%	91/93
Q-learning ( $\rho = 0.9, \epsilon = 0.75$ )	4.41796 $\pm$ 0.1043	1336937.3000 $\pm$ 3633.1147	0.0%	0.0263 $\pm$ 0.0069	7408.0000 $\pm$ 4975.6333	0.4211 $\pm$ 0.0219	132300.0000 $\pm$ 1661.3242	100.0%	91/91
Q-learning ( $\rho = 0.99, \epsilon = 0.75$ )	4.48334 $\pm$ 0.0950	1321232.6000 $\pm$ 4239.8322	0.0%	0.0246 $\pm$ 0.0126	7053.3000 $\pm$ 3504.0261	0.3902 $\pm$ 0.0200	114900.0000 $\pm$ 2507.9872	100.0%	91/91
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	2.81375 $\pm$ 0.6228	807321.3000 $\pm$ 175701.3814	100.0%	0.0539 $\pm$ 0.0272	16241.7000 $\pm$ 8263.9820	0.8684 $\pm$ 0.0333	254400.0000 $\pm$ 7459.2225	100.0%	91/93
Q-learning ( $\rho = 0.7, \epsilon = 0.9$ )	5.23973 $\pm$ 3.3760	1490323.2000 $\pm$ 95458.1602	90.0%	0.0490 $\pm$ 0.0256	14057.6000 $\pm$ 7212.0513	0.7160 $\pm$ 0.0281	204600.0000 $\pm$ 10041.9122	100.0%	91/91
Q-learning ( $\rho = 0.9, \epsilon = 0.9$ )	8.82929 $\pm$ 2.2027	3298690.7000 $\pm$ 713082.8484	100.0%	0.0620 $\pm$ 0.0330	18987.0000 $\pm$ 8803.0574	0.6247 $\pm$ 0.0540	182200.0000 $\pm$ 2398.6993	100.0%	91/91
Q-learning ( $\rho = 0.99, \epsilon = 0.9$ )	12.12238 $\pm$ 0.1392	3727107.0000 $\pm$ 25514.4807	0.0%	0.0660 $\pm$ 0.0335	20051.8000 $\pm$ 10664.1973	0.5289 $\pm$ 0.0282	163800.0000 $\pm$ 2852.1184	100.0%	91/92
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	12.78287 $\pm$ 3.3036	3996847.8000 $\pm$ 100301.9982	100.0%	0.4339 $\pm$ 0.3682	137499.1000 $\pm$ 117908.5560	7.9335 $\pm$ 1.4208	2479600.0000 $\pm$ 452224.3883	100.0%	91/93
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	8.82929 $\pm$ 2.2027	3298690.7000 $\pm$ 713082.8484	100.0%	0.4017 $\pm$ 0.4531	122460.0000 $\pm$ 140043.9498	7.4290 $\pm$ 0.9222	2117200.0000 $\pm$ 289360.0055	100.0%	91/91
Q-learning ( $\rho = 0.9, \epsilon = 1$ )	8.27295 $\pm$ 1.7336	2561796.4000 $\pm$ 52687.9884	100.0%	0.3989 $\pm$ 0.2493	126615.0000 $\pm$ 82858.2111	5.3974 $\pm$ 1.8055	1678400.0000 $\pm$ 563602.3776	100.0%	91/92
Q-learning ( $\rho = 0.99, \epsilon = 1$ )	10.17175 $\pm$ 4.4533	3250319.2000 $\pm$ 146198.1491	100.0%	0.2085 $\pm$ 0.2196	66737.3000 $\pm$ 70059.1983	5.4702 $\pm$ 2.0360	1745200.0000 $\pm$ 650229.7748	100.0%	91/93
Async Value Iteration	0.12174 $\pm$ 0.0075	130220.0000 $\pm$ 0.0000	N/A						

**Table 50.** Comparing performance of Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.999$  (Problem 13). Two versions of value iteration are also appended for contrast

Algorithm (Problem 13) ( $\gamma = 0.999$ )	Runtime (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	0.42100 $\pm$ 0.0143	127752.2000 $\pm$ 108.4627	0.0%	0.0047 $\pm$ 0.0008	1477.1000 $\pm$ 76.6022	0.1265 $\pm$ 0.0068	3910.0000 $\pm$ 300.0000	100.0%	33/33
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	0.38638 $\pm$ 0.0109	119281.5000 $\pm$ 378.1992	0.0%	0.0042 $\pm$ 0.0007	1313.6000 $\pm$ 65.3579	0.0914 $\pm$ 0.0029	2840.0000 $\pm$ 489.8979	100.0%	33/33
Q-learning ( $\rho = 0.9, \epsilon = 0$ )	0.37736 $\pm$ 0.0127	117003.6000 $\pm$ 1313.0820	0.0%	0.0044 $\pm$ 0.0008	1309.9000 $\pm$ 33.8274	0.0727 $\pm$ 0.0038	2280.0000 $\pm$ 600.0000	100.0%	33/35
Q-learning ( $\rho = 0.99, \epsilon = 0$ )	0.36804 $\pm$ 0.0097	115982.0000 $\pm$ 1690.8937	0.0%	0.0042 $\pm$ 0.0007	1369.4000 $\pm$ 153.5579	0.0646 $\pm$ 0.0044	2080.0000 $\pm$ 871.7798	100.0%	33/35
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	0.52299 $\pm$ 0.0136	170583.4000 $\pm$ 373.3465	0.0%	0.0051 $\pm$ 0.0016	1629.0000 $\pm$ 322.1391	0.1338 $\pm$ 0.0029	4290.0000 $\pm$ 300.0000	100.0%	33/33
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	0.50840 $\pm$ 0.0105	161753.3000 $\pm$ 484.6933	0.0%	0.0049 $\pm$ 0.0008	1532.1000 $\pm$ 224.2046	0.1000 $\pm$ 0.0029	3210.0000 $\pm$ 300.0000	100.0%	33/33
Q-learning ( $\rho = 0.9, \epsilon = 0.25$ )	0.49029 $\pm$ 0.0129	156851.3000 $\pm$ 651.1839	0.0%	0.0053 $\pm$ 0.0012	1599.5000 $\pm$ 332.4681	0.0806 $\pm$ 0.0070	2540.0000 $\pm$ 489.8979	100.0%	33/33
Q-learning ( $\rho = 0.99, \epsilon = 0.25$ )	0.48049 $\pm$ 0.0121	155549.2000 $\pm$ 427.4459	0.0%	0.0045 $\pm$ 0.0011	1478.0000 $\pm$ 309.7082	0.0749 $\pm$ 0.0037	2410.0000 $\pm$ 538.5165	100.0%	33/34
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	0.80326 $\pm$ 0.0161	259210.8000 $\pm$ 1509.3220	0.0%	0.0062 $\pm$ 0.0016	2006.1000 $\pm$ 345.7010	0.1541 $\pm$ 0.0074	4970.0000 $\pm$ 640.3124	100.0%	33/33
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	0.77900 $\pm$ 0.0185	252295.5000 $\pm$ 1926.2094	0.0%	0.0068 $\pm$ 0.0020	2069.7000 $\pm$ 648.0762	0.1176 $\pm$ 0.0055	5070.0000 $\pm$ 1791.6423	100.0%	33/33
Q-learning ( $\rho = 0.9, \epsilon = 0.5$ )	0.76413 $\pm$ 0.0141	248537.1000 $\pm$ 1056.6522	0.0%	0.0057 $\pm$ 0.0014	2003.1000 $\pm$ 303.4526	0.0934 $\pm$ 0.0028	3100.0000 $\pm$ 0.0000	100.0%	33/33
Q-learning ( $\rho = 0.99, \epsilon = 0.5$ )	0.77291 $\pm$ 0.0190	247948.2000 $\pm$ 1162.4237	0.0%	0.0061 $\pm$ 0.0019	1859.1000 $\pm$ 547.9541	0.0910 $\pm$ 0.0072	2900.0000 $\pm$ 447.2136	100.0%	33/33
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	1.71760 $\pm$ 0.0251	573371.0000 $\pm$ 6643.6816	0.0%	0.0083 $\pm$ 0.0020	2860.5000 $\pm$ 685.0938	0.1806 $\pm$ 0.0083	6430.0000 $\pm$ 1100.0000	100.0%	33/33
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	1.76432 $\pm$ 0.0306	565352.7000 $\pm$ 5937.4857	0.0%	0.0079 $\pm$ 0.0021	2409.1000 $\pm$ 429.4284	0.1503 $\pm$ 0.0055	5070.0000 $\pm$ 1791.6423	100.0%	33/33
Q-learning ( $\rho = 0.9, \epsilon = 0.75$ )	1.79295 $\pm$ 0.0416	592897.9000 $\pm$ 6914.9360	0.0%	0.0091 $\pm$ 0.0031	2858.0000 $\pm$ 817.7569	0.1283 $\pm$ 0.0073	4260.0000 $\pm$ 800.0000	100.0%	33/33
Q-learning ( $\rho = 0.99, \epsilon = 0.75$ )	1.74500 $\pm$ 0.0173	587144.2000 $\pm$ 4339.3984	0.0%	0.0064 $\pm$ 0.0016	2172.8000 $\pm$ 602.0843	0.1150 $\pm$ 0.0054	4010.0000 $\pm$ 1513.2746	100.0%	33/34
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	0.55173 $\pm$ 0.0594	189588.4000 $\pm$ 17275.8681	100.0%	0.0103 $\pm$ 0.0033	3651.0000 $\pm$ 1959.4914	0.2617 $\pm$ 0.0102	8950.0000 $\pm$ 3008.3218	100.0%	33/33
Q-learning ( $\rho = 0.7, \epsilon = 0.9$ )	0.45832 $\pm$ 0.0534	159215.7000 $\pm$ 17722.4830	100.0%	0.0104 $\pm$ 0.0053	3611.0000 $\pm$ 1702.7189	0.2070 $\pm$ 0.0162	7280.0000 $\pm$ 4354.3082	100.0%	33/33
Q-learning ( $\rho = 0.9, \epsilon = 0.9$ )	0.42473 $\pm$ 0.0747	142218.5000 $\pm$ 25592.6479	100.0%	0.0127 $\pm$ 0.0065	4168.0000 $\pm$ 2153.3371	0.1819 $\pm$ 0.0126	6080.0000 $\pm$ 3280.2439	100.0%	33/33
Q-learning ( $\rho = 0.99, \epsilon = 0.9$ )	0.35403 $\pm$ 0.0764	119643.8000 $\pm$ 25616.2213	100.0%	0.0107 $\pm$ 0.0045	3988.7000 $\pm$ 1632.3342	0.1696 $\pm$ 0.0126	5750.0000 $\pm$ 3383.7849	100.0%	33/33
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	1.44662 $\pm$ 0.2058	409898.4000 $\pm$ 71658.6692	100.0%	0.0146 $\pm$ 0.0141	5338.0000 $\pm$ 5032.0724	0.1786 $\pm$ 0.0124	27500.0000 $\pm$ 44468.9780	100.0%	33/35
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	0.99436 $\pm$ 0.1848	345500.0000 $\pm$ 61741.7089	100.0%	0.0128 $\pm$ 0.0110	4797.6000 $\pm$ 4424.0289	0.1523 $\pm$ 0.0092	22770.0000 $\pm$ 32972.8676	100.0%	33/33
Q-learning ( $\rho = 0.9, \epsilon = 1$ )	0.76700 $\pm$ 0.1771	265314.4000 $\pm$ 69057.6632	100.0%	0.0081 $\pm$ 0.0051	2921.1000 $\pm$ 1713.3067	0.1492 $\pm$ 0.0841	17020.0000 $\pm$ 28242.6372	100.0%	33/33
Q-learning ( $\rho = 0.99, \epsilon = 1$ )	0.62492 $\pm$ 0.1279	218100.4000 $\pm$ 43522.9246	100.0%	0.0106 $\pm$ 0.0065	3915.7000 $\pm$ 2612.0314	0.1477 $\pm$ 0.0718	14540.0000 $\pm$ 26131.2074	100.0%	33/33
Async Value Iteration	0.07626 $\pm$ 0.0034	112896.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	33/33
Value Iteration	0.11517 $\pm$ 0.0051	163968.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	33/33

**Table 51.** Comparing performance of Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.99$  (Problem 14). Two versions of value iteration are also appended for contrast

Algorithm (Problem 14) ( $\gamma = 0.99$ )	Runtime (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path	
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	0.30970 $\pm$ 0.0091	96328.6000 $\pm$ 176.8594	0.0%	0.0010 $\pm$ 0.0000	1252.0000 $\pm$ 265.3127	0.1077 $\pm$ 0.0078	2480.0000 $\pm$ 1333.0303	100.0%	24/22	
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	0.29002 $\pm$ 0.0048	91828.5000 $\pm$ 931.1332	0.0%	0.0049 $\pm$ 0.0014	1435.9000 $\pm$ 234.8917	0.0769 $\pm$ 0.0081	2460.0000 $\pm$ 2107.1308	100.0%	24/28	
Q-learning ( $\rho = 0.9, \epsilon = 0$ )	0.30375 $\pm$ 0.0170	93267.8000 $\pm$ 1501.4346	0.0%	0.0046 $\pm$ 0.0012	1352.2000 $\pm$ 270.5168	0.0645 $\pm$ 0.0080	2000.0000 $\pm$ 2529.8221	50.0%	24/28	
Q-learning ( $\rho = 0.99, \epsilon = 0$ )	0.31794 $\pm$ 0.0239	94536.9000 $\pm$ 1658.8502	0.0%	0.0044 $\pm$ 0.0008	1390.3000 $\pm$ 218.5599	0.1913 $\pm$ 0.0751	5766.6667 $\pm$ 20717.6700	60.0%	24/29	
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	0.40140 $\pm$ 0.0155	126466.8000 $\pm$ 334.4978	0.0%	0.0063 $\pm$ 0.0008	1860.4000 $\pm$ 358.0489	0.1197 $\pm$ 0.0136	3800.0000 $\pm$ 3768.2887	100.0%	24/26	
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	0.38819 $\pm$ 0.0115	120149.8000 $\pm$ 280.1523	0.0%	0.0050 $\pm$ 0.0014	1568.2000 $\pm$ 362.5719	0.1007 $\pm$ 0.0166	3210.0000 $\pm$ 5166.2365	100.0%	24/25	
Q-learning ( $\rho = 0.9, \epsilon = 0.25$ )	0.36089 $\pm$ 0.0113	117124.6000 $\pm$ 603.3306	0.0%	0.0047 $\pm$ 0.0013	1479.7000 $\pm$ 294.5665	0.1410 $\pm$ 0.0746	4630.0000 $\pm$ 25772.2719	100.0%	24/26	
Q-learning ( $\rho = 0.99, \epsilon = 0.25$ )	0.38105 $\pm$ 0.0138	116433.2000 $\pm$ 451.0545	0.0%	0.0063 $\pm$ 0.0016	1825.3000 $\pm$ 231.3846	0.1326 $\pm$ 0.0000	4200.0000 $\pm$ 0.0000	10.0%	24/25	
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	0.59012 $\pm$ 0.0191	198592.8000 $\pm$ 240.0416	0.0%	0.0071 $\pm$ 0.0019	2195.3000 $\pm$ 402.9462	0.1289 $\pm$ 0.0045	4140.0000 $\pm$ 1496.0520	100.0%	24/26	
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	0.56367 $\pm$ 0.0122	184019.0000 $\pm$ 561.0007	0.0%	0.0065 $\pm$ 0.0014	2107.4000 $\pm$ 306.1615	0.1137 $\pm$ 0.0268	3700.0000 $\pm$ 7549.8344	100.0%	24/26	
Q-learning ( $\rho = 0.9, \epsilon = 0.5$ )	0.55118 $\pm$ 0.0081	182080.0000 $\pm$ 618.4603	0.0%	0.0059 $\pm$ 0.0018	1922.5000 $\pm$ 459.6375	0.1083 $\pm$ 0.0255	3570.0000 $\pm$ 8343.2608	100.0%	24/27	
Q-learning ( $\rho = 0.99, \epsilon = 0.5$ )	0.55182 $\pm$ 0.0097	181445.6000 $\pm$ 872.0140	0.0%	0.0065 $\pm$ 0.0014	2058.4000 $\pm$ 397.7271	0.2859 $\pm$ 0.2007	9600.0000 $\pm$ 67000.0000	20.0%	24/24	
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	1.18849 $\pm$ 0.0185	397074.7000 $\pm$ 1116.4341	0.0%	0.0067 $\pm$ 0.0021	2207.1000 $\pm$ 846.5800	0.1594 $\pm$ 0.0065	5430.0000 $\pm$ 2968.1614	100.0%	24/25	
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	1.17601 $\pm$ 0.0220	392153.6000 $\pm$ 2179.3055	0.0%	0.0079 $\pm$ 0.0032	2882.2000 $\pm$ 1006.5925	0.1436 $\pm$ 0.0217	4850.0000 $\pm$ 5852.3500	100.0%	24/25	
Q-learning ( $\rho = 0.9, \epsilon = 0.75$ )	1.19923 $\pm$ 0.0324	390143.3000 $\pm$ 2422.7454	0.0%	0.0067 $\pm$ 0.0021	2280.6000 $\pm$ 621.1773	0.1828 $\pm$ 0.0843	61200.0000 $\pm$ 28999.1103	100.0%	24/26	
Q-learning ( $\rho = 0.99, \epsilon = 0.75$ )	1.16542 $\pm$ 0.0128	394794.1000 $\pm$ 2573.2271	0.0%	0.0088 $\pm$ 0.0036	2861.0000 $\pm$ 829.6588	nan	nan	nan	0.0%	24/26
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	0.89534 $\pm$ 0.0161	104044.2000 $\pm$ 970.2659	0.0%	0.0073 $\pm$ 0.0019	2195.3000 $\pm$ 402.9462	0.1279 $\pm$ 0.0040	8130.0000 $\pm$ 16174.3624	100.0%	24/27	
Q-learning ( $\rho = 0.7, \epsilon = 0.9$ )	0.304500 $\pm$ 0.0401	1042682.5000 $\pm$ 8516.5642	0.0%	0.0061 $\pm$ 0.0034	2100.2000 $\pm$ 1257.2713	0.1948 $\pm$ 0.0451	6610.0000 $\pm$ 13794.5641	100.0%	24/24	
Q-learning ( $\rho = 0.9, \epsilon = 0.9$ )	0.307272 $\pm$ 0.0809	1043268.7000 $\pm$ 13685.6012	0.0%	0.0100 $\pm$ 0.0045	3357.6000 $\pm$ 1384.5272	0.2213 $\pm$ 0.1023	7600.0000 $\pm$ 35547.1518	100.0%	24/26	
Q-learning ( $\rho = 0.99, \epsilon = 0.9$ )	0.315766 $\pm$ 0.0590	1075096.7000 $\pm$ 10863.5693	0.0%	0.0082 $\pm$ 0.0058	2911.9000 $\pm$ 1920.4674	nan	nan	nan	0.0%	24/27
Q-learning ( $\rho = 0.5, \epsilon = 0.99$ )	13.48848 $\pm$ 0.2029	4627904.0000 $\pm$ 42637.9797	0.0%	0.0069 $\pm$ 0.0042	1833.1000 $\pm$ 1449.0801	0.3290 $\pm$ 0.0742	11200.0000 $\pm$ 21090.2821	100.0%	24/25	
Q-learning ( $\rho = 0.7, \epsilon = 0.99$ )	1.17601 $\pm$ 0.0220	392153.6000 $\pm$ 2179.3055	0.0%	0.0079 $\pm$ 0.0032	2882.2000 $\pm$ 1006.5925	0.1436 $\pm$ 0.0217	4850.0000 $\pm$ 5852.3500	100.0%	24/25	
Q-learning ( $\rho = 0.9, \epsilon = 0.99$ )	1.19923 $\pm$ 0.0324	390143.3000 $\pm$ 2422.7454	0.0%	0.0067 $\pm$ 0.0021	2280.6000 $\pm$ 621.1773	0.1828 $\pm$ 0.0843	61200.0000 $\pm$ 28999.1103	100.0%	24/26	
Q-learning ( $\rho = 0.99, \epsilon = 0.99$ )	1.16542 $\pm$ 0.0128	394794.1000 $\pm$ 2573.2271	0.0%	0.0088 $\pm$ 0.0036	2861.0000 $\pm$ 829.6588	nan	nan	nan	0.0%	24/26
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	0.89534 $\pm$ 0.0161	104044.2000 $\pm$ 970.2659	0.0%	0.0073 $\pm$ 0.0019	2195.3000 $\pm$ 402.9462	0.1279 $\pm$ 0.0040	8130.0000 $\pm$ 16174.3624	100.0%	24/27	
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	0.304500 $\pm$ 0.0401	1042682.5000 $\pm$ 8516.5642	0.0%	0.0061 $\pm$ 0.0034	2100.2000 $\pm$ 1257.2713	0.1948 $\pm$ 0.0451	6610.0000 $\pm$			

**Table 53.** Comparing performance of Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.99$  (Problem 15). Two versions of value iteration are also appended for contrast

Algorithm (Problem 15) ( $\gamma = 0.99$ )	Runtime (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	0.2852 $\pm$ 0.0117	80628.3000 $\pm$ 184.7880	0.0%	0.0038 $\pm$ 0.0067	1428.2000 $\pm$ 129.3766	0.1211 $\pm$ 0.0113	3860.0000 $\pm$ 3261.9013	100.0%	22/24
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	0.29617 $\pm$ 0.0100	83249.1000 $\pm$ 321.7780	0.0%	0.0053 $\pm$ 0.0012	1495.4000 $\pm$ 203.8942	0.0940 $\pm$ 0.0080	29900.0000 $\pm$ 2426.9322	100.0%	22/25
Q-learning ( $\rho = 0.9, \epsilon = 0$ )	0.26301 $\pm$ 0.0066	81721.2000 $\pm$ 1224.0630	0.0%	0.0055 $\pm$ 0.0008	1565.3000 $\pm$ 179.0643	0.0876 $\pm$ 0.0136	27000.0000 $\pm$ 4404.5431	100.0%	22/26
Q-learning ( $\rho = 0.99, \epsilon = 0$ )	0.26566 $\pm$ 0.0046	81982.9000 $\pm$ 1051.3302	0.0%	0.0049 $\pm$ 0.0012	1413.1000 $\pm$ 133.2039	0.1278 $\pm$ 0.0369	40000.0000 $\pm$ 11175.3697	90.0%	22/26
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	0.36004 $\pm$ 0.0139	113794.9000 $\pm$ 352.9301	0.0%	0.0058 $\pm$ 0.0018	1701.5000 $\pm$ 144.5973	0.1777 $\pm$ 0.0603	56000.0000 $\pm$ 18487.8338	100.0%	22/24
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	0.34671 $\pm$ 0.0133	107591.1000 $\pm$ 295.7453	0.0%	0.0050 $\pm$ 0.0013	1596.3000 $\pm$ 309.2633	0.2252 $\pm$ 0.0689	70333.3333 $\pm$ 19595.9179	90.0%	22/23
Q-learning ( $\rho = 0.9, \epsilon = 0.25$ )	0.33285 $\pm$ 0.0086	104282.5000 $\pm$ 301.7327	0.0%	0.0052 $\pm$ 0.0010	1704.4000 $\pm$ 263.9326	0.2373 $\pm$ 0.0904	73750.0000 $\pm$ 28145.8256	40.0%	22/23
Q-learning ( $\rho = 0.99, \epsilon = 0.25$ )	0.29281 $\pm$ 0.0066	103099.3000 $\pm$ 399.7337	0.0%	0.0055 $\pm$ 0.0020	1501.7000 $\pm$ 434.1223	nan $\pm$ nan	nan $\pm$ nan	0.0%	22/25
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	0.51373 $\pm$ 0.0124	165942.8000 $\pm$ 657.2284	0.0%	0.0075 $\pm$ 0.0034	2317.4000 $\pm$ 602.3433	0.2117 $\pm$ 0.0821	69200.0000 $\pm$ 26794.7756	100.0%	22/22
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	0.49143 $\pm$ 0.0062	159776.7000 $\pm$ 586.9612	0.0%	0.0055 $\pm$ 0.0015	1763.9000 $\pm$ 409.8842	0.1953 $\pm$ 0.0493	62900.0000 $\pm$ 15712.7337	100.0%	22/22
Q-learning ( $\rho = 0.9, \epsilon = 0.5$ )	0.48423 $\pm$ 0.0164	156980.0000 $\pm$ 740.5661	0.0%	0.0057 $\pm$ 0.0017	1920.3000 $\pm$ 280.7729	0.2244 $\pm$ 0.0650	75250.0000 $\pm$ 23413.4043	40.0%	22/23
Q-learning ( $\rho = 0.99, \epsilon = 0.5$ )	0.48535 $\pm$ 0.0111	156709.0000 $\pm$ 711.7792	0.0%	0.0045 $\pm$ 0.0011	1662.9000 $\pm$ 376.2540	nan $\pm$ nan	nan $\pm$ nan	0.0%	22/23
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	0.89400 $\pm$ 0.0171	329468.7000 $\pm$ 2534.9306	0.0%	0.0076 $\pm$ 0.0021	2541.0000 $\pm$ 668.8504	0.2437 $\pm$ 0.0831	83200.0000 $\pm$ 22076.2316	100.0%	22/25
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	0.98745 $\pm$ 0.0149	326236.2000 $\pm$ 2648.0536	0.0%	0.0059 $\pm$ 0.0022	2039.6000 $\pm$ 651.3889	0.3675 $\pm$ 0.2205	122700.0000 $\pm$ 70218.3025	100.0%	22/25
Q-learning ( $\rho = 0.9, \epsilon = 0.75$ )	0.97015 $\pm$ 0.0175	324972.7000 $\pm$ 1958.8229	0.0%	0.0055 $\pm$ 0.0025	1835.6000 $\pm$ 908.8369	0.4076 $\pm$ 0.1862	137500.0000 $\pm$ 60864.6038	80.0%	22/23
Q-learning ( $\rho = 0.99, \epsilon = 0.75$ )	0.98080 $\pm$ 0.0149	324029.9000 $\pm$ 2793.0568	0.0%	0.0063 $\pm$ 0.0019	2181.8000 $\pm$ 593.4130	nan $\pm$ nan	nan $\pm$ nan	0.0%	22/23
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	2.39904 $\pm$ 0.0339	816977.1000 $\pm$ 8385.0337	0.0%	0.0061 $\pm$ 0.0044	2062.8000 $\pm$ 1377.8644	0.4464 $\pm$ 0.2158	153400.0000 $\pm$ 65871.2067	100.0%	22/23
Q-learning ( $\rho = 0.7, \epsilon = 0.9$ )	2.41486 $\pm$ 0.0513	816467.0000 $\pm$ 6914.5546	0.0%	0.0051 $\pm$ 0.0028	1678.7000 $\pm$ 853.0287	0.6450 $\pm$ 0.5411	219100.0000 $\pm$ 181662.5718	100.0%	22/22
Q-learning ( $\rho = 0.9, \epsilon = 0.9$ )	2.40087 $\pm$ 0.0353	820875.1000 $\pm$ 4842.8374	0.0%	0.0042 $\pm$ 0.0023	1436.8000 $\pm$ 737.5390	1.1962 $\pm$ 0.6682	410222.2222 $\pm$ 229148.5582	90.0%	22/24
Q-learning ( $\rho = 0.99, \epsilon = 0.9$ )	2.48344 $\pm$ 0.0245	840180.2000 $\pm$ 4347.4014	0.0%	0.0053 $\pm$ 0.0022	2036.9000 $\pm$ 1317.5688	nan $\pm$ nan	nan $\pm$ nan	0.0%	22/23
Q-learning ( $\rho = 0.5, \epsilon = 0.99$ )	0.29904 $\pm$ 0.0339	816977.1000 $\pm$ 8385.0337	0.0%	0.0061 $\pm$ 0.0044	2062.8000 $\pm$ 1377.8644	0.4464 $\pm$ 0.2158	153400.0000 $\pm$ 65871.2067	100.0%	22/23
Q-learning ( $\rho = 0.7, \epsilon = 0.99$ )	2.41486 $\pm$ 0.0513	816467.0000 $\pm$ 6914.5546	0.0%	0.0051 $\pm$ 0.0028	1678.7000 $\pm$ 853.0287	0.6450 $\pm$ 0.5411	219100.0000 $\pm$ 181662.5718	100.0%	22/22
Q-learning ( $\rho = 0.9, \epsilon = 0.99$ )	2.40087 $\pm$ 0.0353	820875.1000 $\pm$ 4842.8374	0.0%	0.0042 $\pm$ 0.0023	1436.8000 $\pm$ 737.5390	1.1962 $\pm$ 0.6682	410222.2222 $\pm$ 229148.5582	90.0%	22/24
Q-learning ( $\rho = 0.99, \epsilon = 0.99$ )	2.48344 $\pm$ 0.0245	840180.2000 $\pm$ 4347.4014	0.0%	0.0053 $\pm$ 0.0022	2036.9000 $\pm$ 1317.5688	nan $\pm$ nan	nan $\pm$ nan	0.0%	22/23
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	10.01520 $\pm$ 0.1727	3441286.8000 $\pm$ 42420.9373	0.0%	0.0024 $\pm$ 0.0013	707.3000 $\pm$ 387.3637	0.7493 $\pm$ 0.4824	260100.0000 $\pm$ 167507.2834	100.0%	22/23
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	10.01022 $\pm$ 0.0884	3453719.7000 $\pm$ 35047.7117	0.0%	0.0044 $\pm$ 0.0030	1478.9000 $\pm$ 956.0314	1.3257 $\pm$ 0.9365	464000.0000 $\pm$ 325482.7184	100.0%	22/23
Q-learning ( $\rho = 0.9, \epsilon = 1$ )	9.96116 $\pm$ 0.1227	3429200.0000 $\pm$ 43046.5147	0.0%	0.0042 $\pm$ 0.0037	1381.5000 $\pm$ 1143.9474	1.2725 $\pm$ 1.0050	437700.0000 $\pm$ 344933.9212	100.0%	22/23
Q-learning ( $\rho = 0.99, \epsilon = 1$ )	9.92413 $\pm$ 0.1156	3417329.0000 $\pm$ 41567.4616	0.0%	0.0041 $\pm$ 0.0047	957.2000 $\pm$ 1273.1382	nan $\pm$ nan	nan $\pm$ nan	0.0%	22/23
Async Value Iteration	0.11313 $\pm$ 0.0060	161800.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	22/23
Value Iteration	0.15088 $\pm$ 0.0044	213576.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	22/24

**Table 54.** Comparing performance of Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.999$  (Problem 15). Two versions of value iteration are also appended for contrast

Algorithm (Problem 15) ( $\gamma = 0.999$ )	Runtime (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	0.26376 $\pm$ 0.0114	87332.2000 $\pm$ 130.8012	0.0%	0.0041 $\pm$ 0.0006	1229.7000 $\pm$ 55.2323	0.1200 $\pm$ 0.0058	35800.0000 $\pm$ 400.0000	100.0%	22/23
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	0.25831 $\pm$ 0.0102	80872.9000 $\pm$ 135.5858	0.0%	0.0040 $\pm$ 0.0007	1255.4000 $\pm$ 68.8421	0.0817 $\pm$ 0.0078	25600.0000 $\pm$ 489.8079	100.0%	22/23
Q-learning ( $\rho = 0.9, \epsilon = 0$ )	0.25036 $\pm$ 0.0083	77700.9000 $\pm$ 434.0469	0.0%	0.0038 $\pm$ 0.0009	1217.4000 $\pm$ 47.3861	0.0654 $\pm$ 0.0032	20000.0000 $\pm$ 0.0000	100.0%	22/24
Q-learning ( $\rho = 0.99, \epsilon = 0$ )	0.24895 $\pm$ 0.0086	78801.6000 $\pm$ 144.6556	0.0%	0.0036 $\pm$ 0.0009	1220.8000 $\pm$ 49.0649	0.0562 $\pm$ 0.0015	18000.0000 $\pm$ 0.0000	100.0%	22/22
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	0.35408 $\pm$ 0.0148	111584.0000 $\pm$ 412.9496	0.0%	0.0051 $\pm$ 0.0006	1864.1000 $\pm$ 289.8474	0.1230 $\pm$ 0.0039	39300.0000 $\pm$ 458.2576	100.0%	22/22
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	0.33355 $\pm$ 0.0096	105435.9000 $\pm$ 445.6337	0.0%	0.0046 $\pm$ 0.0009	1547.0000 $\pm$ 217.0396	0.0906 $\pm$ 0.0045	28700.0000 $\pm$ 458.2576	100.0%	22/22
Q-learning ( $\rho = 0.9, \epsilon = 0.25$ )	0.32478 $\pm$ 0.0112	102077.8000 $\pm$ 445.8764	0.0%	0.0052 $\pm$ 0.0010	1663.0000 $\pm$ 226.2830	0.0724 $\pm$ 0.0037	23900.0000 $\pm$ 0.0000	100.0%	22/22
Q-learning ( $\rho = 0.99, \epsilon = 0.25$ )	0.32277 $\pm$ 0.0091	102165.3000 $\pm$ 280.8591	0.0%	0.0051 $\pm$ 0.0013	1641.7000 $\pm$ 245.3341	0.0625 $\pm$ 0.0023	20300.0000 $\pm$ 458.2576	100.0%	22/22
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	0.51447 $\pm$ 0.0194	161853.1000 $\pm$ 806.6378	0.0%	0.0066 $\pm$ 0.0014	1961.0000 $\pm$ 305.9297	0.1420 $\pm$ 0.0061	46900.0000 $\pm$ 458.2576	100.0%	22/24
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	0.48531 $\pm$ 0.0082	156059.0000 $\pm$ 1501.6935	0.0%	0.0058 $\pm$ 0.0021	2013.1000 $\pm$ 561.3414	0.1063 $\pm$ 0.0051	34400.0000 $\pm$ 489.8079	100.0%	22/22
Q-learning ( $\rho = 0.9, \epsilon = 0.5$ )	0.47548 $\pm$ 0.0124	158488.8000 $\pm$ 1231.8054	0.0%	0.0059 $\pm$ 0.0017	1840.0000 $\pm$ 576.0533	0.0844 $\pm$ 0.0030	28000.0000 $\pm$ 632.4555	100.0%	22/22
Q-learning ( $\rho = 0.99, \epsilon = 0.5$ )	0.49243 $\pm$ 0.0149	158697.0000 $\pm$ 1032.6458	0.0%	0.0071 $\pm$ 0.0023	2187.3000 $\pm$ 527.6387	0.0792 $\pm$ 0.0069	25000.0000 $\pm$ 447.2136	100.0%	22/24
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	0.98622 $\pm$ 0.0094	328269.0000 $\pm$ 2378.9721	0.0%	0.0062 $\pm$ 0.0028	2125.7000 $\pm$ 859.0931	0.1737 $\pm$ 0.0042	59300.0000 $\pm$ 781.0250	100.0%	22/22
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	0.97029 $\pm$ 0.0225	322805.7000 $\pm$ 3815.3383	0.0%	0.0065 $\pm$ 0.0037	2080.7000 $\pm$ 1144.6155	0.1348 $\pm$ 0.0066	45800.0000 $\pm$ 690.0000	100.0%	22/22
Q-learning ( $\rho = 0.9, \epsilon = 0.75$ )	1.00538 $\pm$ 0.0284	331516.4000 $\pm$ 2783.8548	0.0%	0.0070 $\pm$ 0.0021	2359.3000 $\pm$ 713.9961	0.1137 $\pm$ 0.0060	38100.0000 $\pm$ 700.0000	100.0%	22/22
Q-learning ( $\rho = 0.99, \epsilon = 0.75$ )	1.03605 $\pm$ 0.0168	345653.4000 $\pm$ 2173.5444	0.0%	0.0075 $\pm$ 0.0020	2409.6000 $\pm$ 652.0138	0.1027 $\pm$ 0.0073	35000.0000 $\pm$ 1612.4515	100.0%	22/22
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	2.41348 $\pm$ 0.0200	823790.2000 $\pm$ 6985.4138	0.0%	0.0057 $\pm$ 0.0024	1960.4000 $\pm$ 799.8494	0.2159 $\pm$ 0.0126	74400.0000 $\pm$ 2576.8197	100.0%	22/22
Q-learning ( $\rho = 0.7, \epsilon = 0.9$ )	2.32525 $\pm$ 0.0329	790112.9000 $\pm$ 10358.2303	0.0%	0.0064 $\pm$ 0.0038	2215.6000 $\pm$ 1445.1619	0.1063 $\pm$ 0.0051	59900.0000 $\pm$ 3112.3765	100.0%	22/22
Q-learning ( $\rho = 0.9, \epsilon = 0.9$ )	2.41103 $\pm$ 0.3058	810447.7000 $\pm$ 104275.5428	40.0%	0.0061 $\pm$ 0.0035	2096.1000 $\pm$ 1082.7082	0.1480 $\pm$ 0.0119	50000.0000 $\pm$ 2366.4319	100.0%	22/22
Q-learning ( $\rho = 0.99, \epsilon = 0.9$ )	2.33104 $\pm$ 0.3619	780753.1000 $\pm$ 126706.0884	50.0%	0.0059 $\pm$ 0.0040	1854.8000 $\pm$ 1323.8872	0.1392 $\pm$ 0.0100	47500.0000 $\pm$ 2459.6748	100.0%	22/24
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	0.65477 $\pm$ 0.0012	227721.9000 $\pm$ 28488.3753	100.0%	0.0038 $\pm$ 0.0021	1269.4000 $\pm$ 920.4766	0.3416 $\pm$ 0.0252	118900.0000 $\pm$ 7777.5317	100.0%	22/22
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	0.65119 $\pm$ 0.0059	172374.4000 $\pm$ 27628.8363	100.0%	0.0039 $\pm$ 0.0042	1192.7000 $\pm$ 992.5647	0.2737 $\pm$ 0.0198	96600.0000 $\pm$ 7989.9937	100.0%	22

**Table 56.** Comparing performance of Q-learning as we change the learning rate  $\rho$  in stochastic problem with  $\gamma = 0.999$  (Problem 16). Two versions of value iteration are also appended for contrast

Algorithm (Problem 16) ( $\gamma = 0.999$ )	Runtime (mean $\pm$ std)	#actions (mean $\pm$ std)	Convergence	Discover goal time (mean $\pm$ std)	Discover goal #actions (mean $\pm$ std)	Optimal Initial Cost2Go Time (mean $\pm$ std)	Optimal Initial Cost2Go #actions (mean $\pm$ std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ( $\rho = 0.5, \epsilon = 0$ )	0.03311 $\pm$ 0.0011	12061.5000 $\pm$ 4.5880	0.0%	0.0010 $\pm$ 0.0000	14.0000 $\pm$ 0.0000	0.0025 $\pm$ 0.0007	1000.0000 $\pm$ 0.0000	100.0%	5/5
Q-learning ( $\rho = 0.7, \epsilon = 0$ )	0.03330 $\pm$ 0.0014	12049.9000 $\pm$ 4.0608	0.0%	0.0010 $\pm$ 0.0000	14.0000 $\pm$ 0.0000	0.0026 $\pm$ 0.0007	1000.0000 $\pm$ 0.0000	100.0%	5/5
Q-learning ( $\rho = 0.9, \epsilon = 0$ )	0.03253 $\pm$ 0.0021	12052.6000 $\pm$ 5.6604	0.0%	nan $\pm$ nan	14.0000 $\pm$ 0.0000	0.0032 $\pm$ 0.0010	1000.0000 $\pm$ 0.0000	100.0%	5/5
Q-learning ( $\rho = 0.99, \epsilon = 0$ )	0.01518 $\pm$ 0.0180	4317.2000 $\pm$ 5066.6709	70.0%	nan $\pm$ nan	14.0000 $\pm$ 0.0000	0.0036 $\pm$ 0.0009	1000.0000 $\pm$ 0.0000	100.0%	5/5
Q-learning ( $\rho = 0.5, \epsilon = 0.25$ )	0.00305 $\pm$ 0.0005	1001.7000 $\pm$ 1.3454	100.0%	0.0010 $\pm$ 0.0000	14.8000 $\pm$ 2.2271	0.0030 $\pm$ 0.0005	1000.0000 $\pm$ 0.0000	100.0%	5/5
Q-learning ( $\rho = 0.7, \epsilon = 0.25$ )	0.00296 $\pm$ 0.0005	1002.9000 $\pm$ 1.6401	100.0%	0.0011 $\pm$ 0.0001	15.8000 $\pm$ 2.6000	0.0030 $\pm$ 0.0005	1000.0000 $\pm$ 0.0000	100.0%	5/5
Q-learning ( $\rho = 0.9, \epsilon = 0.25$ )	0.00316 $\pm$ 0.0003	1002.9000 $\pm$ 2.5080	100.0%	nan $\pm$ nan	16.2000 $\pm$ 2.0881	0.0032 $\pm$ 0.0003	1000.0000 $\pm$ 0.0000	100.0%	5/5
Q-learning ( $\rho = 0.99, \epsilon = 0.25$ )	0.00303 $\pm$ 0.0004	1002.0000 $\pm$ 1.4832	100.0%	nan $\pm$ nan	15.4000 $\pm$ 2.0100	0.0030 $\pm$ 0.0004	1000.0000 $\pm$ 0.0000	100.0%	5/5
Q-learning ( $\rho = 0.5, \epsilon = 0.5$ )	0.00330 $\pm$ 0.0005	1005.7000 $\pm$ 3.3481	100.0%	nan $\pm$ nan	16.2000 $\pm$ 5.4000	0.0033 $\pm$ 0.0005	1000.0000 $\pm$ 0.0000	100.0%	5/7
Q-learning ( $\rho = 0.7, \epsilon = 0.5$ )	0.00352 $\pm$ 0.0012	1104.6000 $\pm$ 303.8109	100.0%	nan $\pm$ nan	20.2000 $\pm$ 5.7585	0.0032 $\pm$ 0.0009	1000.0000 $\pm$ 0.0000	100.0%	5/5
Q-learning ( $\rho = 0.9, \epsilon = 0.5$ )	0.00316 $\pm$ 0.0006	1002.5000 $\pm$ 2.2023	100.0%	nan $\pm$ nan	16.0000 $\pm$ 3.7947	0.0032 $\pm$ 0.0006	1000.0000 $\pm$ 0.0000	100.0%	5/5
Q-learning ( $\rho = 0.99, \epsilon = 0.5$ )	0.00303 $\pm$ 0.0011	1007.1000 $\pm$ 3.5903	100.0%	0.0011 $\pm$ 0.0000	21.4000 $\pm$ 5.4443	0.0030 $\pm$ 0.0011	1000.0000 $\pm$ 0.0000	100.0%	5/5
Q-learning ( $\rho = 0.5, \epsilon = 0.75$ )	0.00305 $\pm$ 0.0001	1008.3000 $\pm$ 6.6340	100.0%	0.0010 $\pm$ 0.0000	25.0000 $\pm$ 13.7186	0.0031 $\pm$ 0.0001	1000.0000 $\pm$ 0.0000	100.0%	5/5
Q-learning ( $\rho = 0.7, \epsilon = 0.75$ )	0.00367 $\pm$ 0.0014	1009.0000 $\pm$ 7.9246	100.0%	nan $\pm$ nan	20.4000 $\pm$ 11.8254	0.0036 $\pm$ 0.0014	1000.0000 $\pm$ 0.0000	100.0%	5/5
Q-learning ( $\rho = 0.9, \epsilon = 0.75$ )	0.00370 $\pm$ 0.0008	1007.3000 $\pm$ 5.5687	100.0%	0.0010 $\pm$ 0.0000	15.6000 $\pm$ 8.2849	0.0036 $\pm$ 0.0008	1000.0000 $\pm$ 0.0000	100.0%	5/5
Q-learning ( $\rho = 0.99, \epsilon = 0.75$ )	0.00354 $\pm$ 0.0012	1104.9000 $\pm$ 299.0797	100.0%	0.0010 $\pm$ 0.0000	21.2000 $\pm$ 9.6830	0.0031 $\pm$ 0.0003	1000.0000 $\pm$ 0.0000	100.0%	5/5
Q-learning ( $\rho = 0.5, \epsilon = 0.9$ )	0.00332 $\pm$ 0.0005	1012.6000 $\pm$ 8.7430	100.0%	nan $\pm$ nan	18.7000 $\pm$ 9.1657	0.0033 $\pm$ 0.0005	1000.0000 $\pm$ 0.0000	100.0%	5/5
Q-learning ( $\rho = 0.7, \epsilon = 0.9$ )	0.00271 $\pm$ 0.0004	1010.9000 $\pm$ 9.3429	100.0%	0.0010 $\pm$ 0.0000	23.6000 $\pm$ 10.6883	0.0026 $\pm$ 0.0004	1000.0000 $\pm$ 0.0000	100.0%	5/5
Q-learning ( $\rho = 0.9, \epsilon = 0.9$ )	0.00337 $\pm$ 0.0007	1016.5000 $\pm$ 12.4197	100.0%	nan $\pm$ nan	20.0000 $\pm$ 7.3212	0.0034 $\pm$ 0.0007	1000.0000 $\pm$ 0.0000	100.0%	5/5
Q-learning ( $\rho = 0.99, \epsilon = 0.9$ )	0.00326 $\pm$ 0.0007	1110.8000 $\pm$ 296.5181	100.0%	0.0010 $\pm$ 0.0000	16.6000 $\pm$ 6.6963	0.0030 $\pm$ 0.0005	1000.0000 $\pm$ 0.0000	100.0%	5/5
Q-learning ( $\rho = 0.5, \epsilon = 1$ )	0.00303 $\pm$ 0.0007	1022.8000 $\pm$ 16.5638	100.0%	nan $\pm$ nan	19.4000 $\pm$ 22.7517	0.0029 $\pm$ 0.0005	1000.0000 $\pm$ 0.0000	100.0%	5/5
Q-learning ( $\rho = 0.7, \epsilon = 1$ )	0.00290 $\pm$ 0.0003	1021.5000 $\pm$ 16.7108	100.0%	0.0010 $\pm$ 0.0000	35.3000 $\pm$ 37.5341	0.0028 $\pm$ 0.0004	1000.0000 $\pm$ 0.0000	100.0%	5/5
Q-learning ( $\rho = 0.9, \epsilon = 1$ )	0.00271 $\pm$ 0.0005	1017.1000 $\pm$ 14.4599	100.0%	nan $\pm$ nan	18.4000 $\pm$ 7.3648	0.0026 $\pm$ 0.0005	1000.0000 $\pm$ 0.0000	100.0%	5/5
Q-learning ( $\rho = 0.99, \epsilon = 1$ )	0.00275 $\pm$ 0.0005	1024.8000 $\pm$ 29.6968	100.0%	0.0010 $\pm$ 0.0000	19.6000 $\pm$ 13.1697	0.0026 $\pm$ 0.0005	1000.0000 $\pm$ 0.0000	100.0%	5/5
Async Value Iteration	0.00030 $\pm$ 0.0005	144.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	5/5
Value Iteration	0.00061 $\pm$ 0.0005	224.0000 $\pm$ 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	5/5