Relating Reinforcement Learning to Dynamic Programming-Based Planning

Filip V. Georgiev[1], Kalle G. Timperi[1], Başak Sakçak[1,2], Steven M. LaValle[1⋆]

[1] Center for Applied Computing
Faculty of Information Technology and Electrical Engineering
University of Oulu, Finland
[2] Dept. of Advanced Computing Sciences, Maastricht University, the Netherlands

Abstract. This paper bridges some of the gap between optimal planning and reinforcement learning (RL), both of which share roots in dynamic programming applied to sequential decision making or optimal control. Whereas planning typically favors deterministic models, goal termination, and cost minimization, RL tends to favor stochastic models, infinite-horizon discounting, and reward maximization in addition to learning-related parameters such as the learning rate and greediness factor. A derandomized version of RL is developed, analyzed, and implemented to yield performance comparisons with value iteration and Dijkstra's algorithm using simple planning models. Next, mathematical analysis shows: 1) conditions under which cost minimization and reward maximization are equivalent, 2) conditions for equivalence of single-shot goal termination and infinite-horizon episodic learning, and 3) conditions under which discounting causes goal achievement to fail. The paper then advocates for defining and optimizing *truecost*, rather than inserting arbitrary parameters to guide operations. Performance studies are extended to the stochastic case, using planning-oriented criteria and comparing value iteration to RL with learning rates and greediness factors.

Keywords: planning algorithms, motion planning, discrete planning, reinforcement learning, Q-learning, value iteration, dynamic programming

1 Introduction

Bellman introduced the principle of dynamic programming in the 1950s as a way to efficiently compute solutions to sequential decision-making problems [3]. The core idea is the *principle of optimality*, a simple observation that optimal solutions must be composed of optimal parts, thereby leading to a powerful recurrence relation or differential equation that underlies the Hamilton-Jacobi-Bellman (HJB) equation in continuous-time optimal control, value and policy

⋆ This work was supported by a European Research Council Advanced Grant (ERC AdG, ILLUSIVE: Foundations of Perception Engineering, 101020977, Academy of Finland BANG! 363637). (e-mail: firstname.lastname@oulu.fi).

iteration methods for discrete-time problems, and is useful for algorithm design and complexity analysis well beyond sequential decision making. In the context of motion planning, Dijkstra's algorithm, A*-search, D* [13,24], and many others compute optimal plans over grids or more general graphs. Dijkstra-like planning algorithms, built upon value iteration, compute approximately optimal feedback plans over continuous spaces using sampling and interpolation of value functions [29]. All of these methods extend gracefully to settings with either nondeterministic (possibilistic) or stochastic (probabilistic) prediction uncertainty [14].

In the 1990s, Barto and Sutton introduced a powerful paradigm that cleverly interleaves the discovery of the effect of actions with the calculations of optimal values and plans [25]. In preceding works, experimental data was first used to learn or estimate a model of the form $x' = f(x, u)$ for states x, x' and action u (or $p(x' \mid x, u)$ in a stochastic setting), and then the model was used as a module in dynamic programming methods. The new paradigm, called *reinforcement learning (RL)*, retains all of the structure of Bellman's original dynamic programming equations, while inserting learning or discovery directly into the algorithms for computing optimal plans (or policies). Thus, RL is a fusion of optimal sequential decision-making (control, planning) with machine learning (system estimation, identification).

In recent years, RL has become ubiquitous as an approach to problems in robotics and embodied AI [12,26]. As it is applied to motion planning, it has been clearly useful as a way to generate motion/control primitives to complement planners (e.g., [7]); however, the relationships across the Bellman-inspired family of methods have become increasingly opaque. In non-RL works, a clear cost model is formulated in terms of engineering resources, such as time or energy expended, and plans are selected to globally optimize cumulative cost. In standard RL, a reward model is defined that is biologically inspired and is often shaped or tuned to yield desired outcomes [19]. Furthermore, RL tends to be formulated as an infinite-horizon problem in which the most preferred way to make cumulative rewards finite is via an arbitrary discount factor that has little meaning in an engineering context (notable exceptions exist [18]). In non-RL approaches to planning, the episode usually ends when goals are achieved; even though the number of steps may be unbounded, it is finite and can be efficiently handled by termination actions [14] or terminal states [6,20], thereby preserving the true cost model. Finally, RL is formulated exclusively in a stochastic setting, whereas for non-RL methods there is a simple and clear relationship between corresponding algorithms in deterministic and stochastic settings. Probabilistic modeling is not the 'only' or 'most general' way to handle uncertainties (see, for example, [11]).

To address these concerns, this paper investigates a series of methods along the spectrum from deterministic planning to stochastic RL (Q-learning). Relationships are established through both theoretical analysis and experimental studies. Section 2 starts with optimal deterministic planning and develops a comparable derandomized analog of RL that clarifies the issues of whether a model is given and whether one can jump at will between arbitrary states. Computation

and convergence rates are compared over several examples. Section 3 bridges the gap between typical planning and RL formulations through rigorous mathematical analysis of their relationship, with an emphasis on using physically motivated costs in models, as opposed to arbitrary discounts and rewards. Section 4 extends the studies to stochastic state-transition models, which are prevalent for RL. Section 5 summarizes the main conclusions and remaining challenges.

2 Optimal Planning and Deterministic RL

2.1 Optimal planning concepts

Consider a standard optimal planning problem expressed as a six-tuple $\mathcal{P} = (X, U, f, x_1, X_G, \ell)$, in which X is a finite *state space*, U is a finite *action space*, $f : X \times U \to X$ is the *state transition function*, $x_I \in X$ is the *initial state*, $X_G \subset X$ is the *goal set*, and $\ell : X \times U \to [0, \infty)$ is the *cost function*. Actually, f is allowed to be a partial function, in which $U(x) \subseteq U$ denotes the actions available from each $x \in X$. A *feedback plan, or policy* is a mapping $\pi : X \to U$, and we use the shorthand $f_\pi(x) := f(x, \pi(x))$ to indicate the next state obtained by applying policy π at state x. A policy π is evaluated from any $x_1 \in X$ using the stage-additive *cost functional*

$$L(\pi, x_1) = \sum_{k=1}^{\infty} l(x_k, \pi(x_k)), \tag{1}$$

in which $k \in \mathbb{N}$ indexes over *stages*, and $x_{k+1} = f_\pi(x_k) = f(x_k, \pi(x_k))$. To take X_G into account, a special *termination action* $u_T \in U(x_g)$ exists from every $x_g \in X_G$, and when applied at stage k, then $x_{i+1} = x_i = x_g$ for all all $i \geq k$ and $l(x_g, u_T) = 0$. This ensures that (1) is finite if the goal is reachable from x_1, in spite of having an infinite horizon, that is, infinite number of stages. A plan or policy π^* is *optimal* if for all other π and all $x_1 \in X$, $L(\pi^*, x_1) \leq L(\pi, x_1)$. Let $G_\pi : X \to [0, \infty)$ be the *cost-to-go* function for a plan π, in which each $G_\pi(x)$ is the total cost obtained in (1) by applying $\pi(x)$ from state $x_1 = x$. Let $G^* = G_{\pi^*}$ be the *optimal cost to go*. Determining G^* and π^* are closely related because:

$$\pi^*(x) \in \arg\min_{u \in U(x)} \{\ell(x, u) + G^*(x')\}, \tag{2}$$

in which $x' = f(x, u)$ (see [14], Ch. 2). In the case of classical planning, the model is known; thus, π^* and G^* can be computed using value iteration or by Dijkstra's algorithm, in which the latter can be seen an efficient version of value iteration. Whereas Dijsktra's algorithm is typically faster than value iteration, value iteration generalizes more easily to stochastic planning and RL. Value iteration proceeds by successively computing cost-to-go functions for each stage, proceeding backwards through time, until the values stabilize (they do not change). For each $x_k \in X$, the value is calculated as

$$G_k^*(x_k) = \min_{u_k \in U(x_k)} \Big\{ l(x_k, u_k) + G_{k+1}^*(x_{k+1}) \Big\}, \tag{3}$$

in which $x_{k+1} = f(x_k, u_k)$. Most RL algorithms essentially create an asynchronous, step-by-step version of (3), in which each step processes a state-action pair (x_k, u_k).

2.2 Handling imperfect models by derandomizing RL

Whereas most RL algorithms are designed considering imperfect models, planning algorithms typically assume a complete model. Therefore, our first task is to carefully relate and compare these. Two critical options emerge: 1) Is the model given to the algorithm? Suppose that in the *model-based* case, $\mathcal{P}$ is completely given to the algorithm. In the *model-free* case, $\mathcal{P}$ is known 'to the gods' but not the algorithm. The robot can only try actions and then sense the next state and incurred cost. 2) Can the robot freely 'jump' between states for planning purposes? *Virtual jumping* occurs in planning by using the model: The algorithm deduces what would happen if action u were applied from state x. By default, we focus on *physical jumping*, in which the robot is moved to a new state.

There are three cases based on the critical options: 1) **Model-based.** If the model is available, then, virtual jumping can be used, allowing algorithms such as Dijkstra or value iteration to solve the optimal planning problem before the robot is ever moved. 2) **Model-free with jumping**. If jumping is free and X were known, then the model can be completely discovered by trying every $u \in U(x)$ from every $x \in X$. Suppose that even X is unknown. In this case, we assume sensing is sufficient so that states can be uniquely labeled as they are discovered, and returns to previous states are correctly determined. Even in this extreme scenario, Dijkstra or value iteration could be applied to derive the optimal plan. The difference compared to the first case is that here, jumping must be physical. (If a cost is assigned for physical jumping, then an intriguing new optimization problem emerges.) 3) **Model-free without jumping.** This case most closely matches common RL formulations. The robot must move along one long path to discover the optimal plan. (The directed transition graph underlying $\mathcal{P}$ is required to have every state reachable from every other.)

Note that since the system is deterministic, every state-action pair needs to be considered only once to completely discover f; however, revisits may be needed to fully compute G^*. Furthermore, f can be represented as a directed multigraph with vertices X and labeled edges $U(x)$ for all $x \in X$. This leads to a simple two-phase algorithm: 1) First, 'physically' explore enough to discover the entire graph (jumping not allowed!); several efficient algorithms exist [1,8]. 2) Then, run Dijkstra's algorithm to calculate G^* and π^*. This, however, does not follow the spirit of most RL, which interleaves learning and optimal planning during the entire execution. To move closer to this approach, we introduce a derandomized version of Q-learning, which serves as a gateway between typical planning and RL scenarios.

Let $Q_\pi : X \times U \to \mathbb{R}$ be the *Q-value function*, obtained from (1) by applying u_1 at x_1, and then applying actions $u_i = \pi(x_i)$ for all $i > 1$. Let $Q^* = Q_{\pi^*}$ be the *optimal Q-value function*. Note that G^* and $u = \pi^*(x)$ can be computed

from Q^* by

$$G^*(x) = \min_{u \in U(x)} \Big\{ Q^*(x,u) \Big\}. \tag{4}$$

Q-learning estimates Q^* through the process of stochastic iteration [25,21]. Using a stochastic state transition model $p(x_{k+1}|x_k, u_k)$ (addressed in Section 4), estimates of Q^*, denoted by $\hat{Q}^*$, are determined iteratively as (see [14], Sec. 10.4):

$$\hat{Q}^*(x,u) \leftarrow (1-\rho)\hat{Q}^*(x,u) + \rho \left(l(x,u) + \min_{u' \in U(x')} \Big\{ \hat{Q}^*(x',u') \Big\} \right), \tag{5}$$

in which x' is obtained by applying u from x in a single step, $\rho \in (0,1)$ is the *learning rate*, and $\leftarrow$ denotes an assignment (not a fixed point equation). Note that jumping is not required: The minimization over all $u' \in U(x')$ involves looking up internally stored Q-values. The parameter ρ is chosen to be smaller if the amount of uncertainty or instability is higher. If every reachable state-action pair is visited infinitely often, then $\hat{Q}^*$ converges to Q^* in probability [5,28].

In the limiting case of a deterministic system, there is no uncertainty, which would suggest setting $\rho = 1$ to obtain our proposed *derandomized Q-learning*:

$$\hat{Q}^*(x,u) \leftarrow l(x,u) + \min_{u' \in U(x')} \Big\{ \hat{Q}^*(x',u') \Big\}. \tag{6}$$

Q-learning is considered *off-policy* in the sense that no particular plan of movement for exploration is assumed, as long as every state-action pair is visited infinitely often. In practice, however, a movement plan is needed. To ensure all state-action pairs are visited infinitely often, two *pure exploration* plans are considered: 1) apply random actions, leading to probabilistic convergence, and 2) apply a universal plan from [27] to obtain deterministic convergence.

Proposition 1. *If every state-action pair (x,u) is visited infinitely often for all $x \in X$ and $u \in U(x)$, and (6) is applied in every step, then after a finite number of iterations, $\hat{Q}^* = Q^*$.*

Proof. Let $Z = X \times U$ be an augmented state space. Let $f_Z : X \times U \times U \to X \times U$ be the state transition function in this new domain defined as $f_Z(x,u,u') = (f(x,u), u')$, in which f is the state transition function for the original problem. Then, (6) can be written in this new augmented space as

$$\hat{Q}^*(z) \leftarrow \ell(z) + \min_{u' \in U(f(x,u))} \{ \hat{Q}^*(z') \}, \tag{7}$$

in which $z' = f_Z(z,u')$. Notice that this is one step of value iteration over the new state space Z, the cost $\ell(z)$ is outside of minimization since the minimization is over u'. Because in Q-learning the values are updated when x is visited and u is applied, (7) can be seen as a step in an asynchronous value iteration with single state updates in contrast with value iteration which does a complete sweep of all the states. Then, the finite time convergence to the optimal values follows from the finite time convergence of asynchronous value iteration for deterministic shortest-path problems [4]. □

As is common in RL, we balance exploration vs. exploitation in the form of an ϵ-*greedy* plan/policy. For some fixed value $\epsilon \in [0, 1]$, the policy applies the next action in the pure exploration plan with probability ϵ; otherwise, it applies the minimizing u' in (6). Thus, it is 'greedy' (exploitation) $1 - \epsilon$ of the time on average.

We follow another common RL practice, which is to allow periodic jumping back to the initial state. Thus, the execution is divided into *episodes*, in which each one corresponds to applying an action sequence from the initial state until either X_G or a maximum iteration limit is reached (to escape greediness-induced traps). This issue does not arise in value iteration because the state becomes frozen once X_G is reached and u_T is applied. The relationship between these two settings is analyzed in Section 3.3.

2.3 Computational comparisons

We compared the methods over several planning problems encoded as grids with standard 4-neighbors. Higher dimensions and continuous problems are avoided here because our aim is to improve understanding and we believe the key issues identified in our studies extend to higher dimensional lattices [15] and continuous problems involving dynamic programming with interpolation [16].

The experiments were conducted using Python[3]. They were run on a desktop computer with a Ryzen 7 3700X CPU (3.6GHz)[4]. Figure 1 shows a representative sample of the problems we investigated; many more results appear in the appendix.

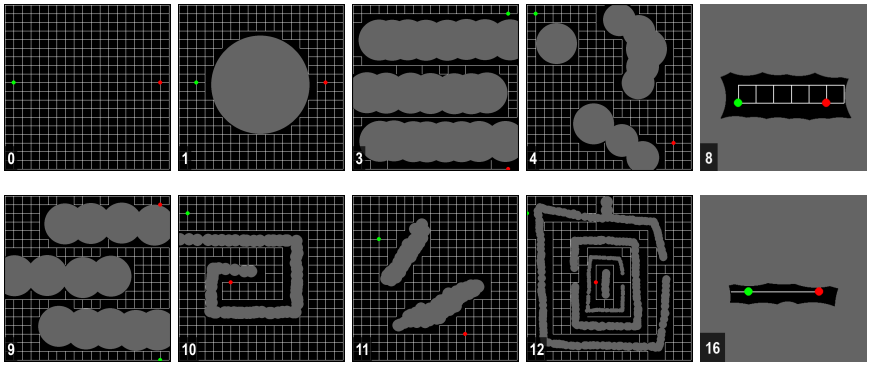

Fig. 1. Ten representative problems (enumerated from top-left to bottom-right as 0, 1, 3, 4, 8, 9, 10, 11, 12, 16). The green dot is the initial state. The red dot is the goal state. The white segments connect the states and can be viewed as the actions allowed by the robot. All actions have the same step cost. The gray areas are obstacles.

[3] The codebase can be found here.

[4] A computer with an i5-12400F CPU (2.50GHz) was used for some of the stochastic experiments ($\gamma \leq 0.9$).

Run times are shown, which closely correlate with number of actions taken (these appear in the appendix). We ran each algorithm 100 times and report the mean and standard deviation. Unless otherwise stated, the number of episodes used for Q-learning is set to 1000, and the number of steps per episode is set to 3000 to be sufficiently longer than any optimal path. These were chosen through trial and error to optimize Q-learning performance. Each run was terminated either after convergence or an iteration limit was reached. We considered three questions:

1. Was a path from the initial to the goal found?
2. Was the path found optimal (is the cost-to-go value from the initial state optimal)?
3. Is the cost-to-go value for each state in the state space optimal?

The first question typically corresponds to a pure planning solution: An action sequence that leads to the goal. Figure 2 shows how varying ϵ affects the time in which the goal state is first discovered. When we initialize all state-action pairs (except the goal) to the same value, then a greedy action would be chosen arbitrarily (depending on which action is first in its enumeration) because all neighboring values are equal. Later, as the cost of any action that has been chosen in previous iterations increases relative to the other available actions the algorithm would prefer unexplored actions. This happens until the goal is discovered and better values are propagated backwards, causing progress-making actions to be preferred. In problems in which there are less states and more obstacles (e.g., Problem 12), the goal is discovered more quickly with a greedier policy. For Problem 0, since the whole state space is available and the goal is relatively far away, a purely greedy approach ($\epsilon = 0$) takes more time to reach it. However, a random policy ($\epsilon = 1$) is quicker because the robot is not hindered by any obstacles.

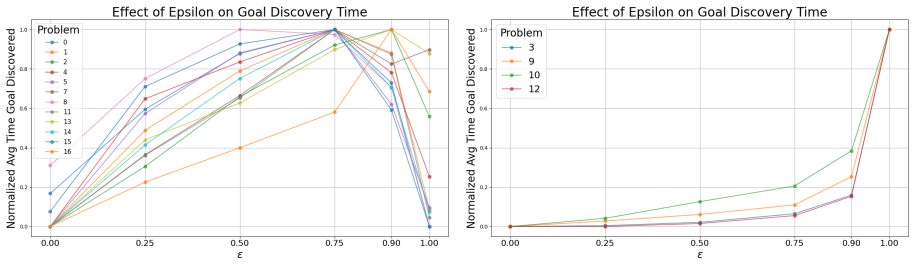

Fig. 2. The effect of decreasing greediness in a policy when applied on different problems.

To address questions (2) and (3) above, we define convergence for Q-learning as convergence to the Q-values consistent with the optimal cost-to-go function. Note that Q^* can be obtained from G^* using (4) for this comparison.

Table 1 shows the performance of the model-free dynamic programming algorithms and that of Q-learning with varying ϵ. We also include a deterministic version of Q-learning, labeled as '(π)', that uses digits of π base four, to select actions [27]. We show the average run time of the algorithm, how often the values for the whole state space converge, and the time it takes for the value of the initial state to converge.

Table 1. Performance of Q-learning and model-free approaches on a deterministic problem (Problem 10). Value iteration shortened to VI.

Algorithm (Problem 10)	Run time (mean ± std)	Convergence	Optimal Initial Cost-to-Go Time (mean ± std)
Q-learning ($\epsilon = 0$)	0.33383 ± 0.0025	0.0 %	0.1720 ± 0.0016
Q-learning ($\epsilon = 0.25$)	0.44088 ± 0.0030	0.0 %	0.2043 ± 0.0041
Q-learning ($\epsilon = 0.5$)	0.64394 ± 0.0045	0.0 %	0.2517 ± 0.0066
Q-learning ($\epsilon = 0.75$)	1.08726 ± 0.2099	76.0 %	0.3502 ± 0.0129
Q-learning ($\epsilon = 0.9$)	0.61889 ± 0.0559	100.0 %	0.4858 ± 0.0239
Q-learning ($\epsilon = 1$)	1.88356 ± 0.6740	100.0 %	1.2836 ± 0.3801
Q-learning (π) ($\epsilon = 1$)	2.02001 ± 0.4996	100.0 %	1.3425 ± 0.2770
Model-free Dijkstra	0.00248 ± 0.0006	N/A	N/A
Model-free Asynchronous VI	0.04920 ± 0.0004	N/A	N/A
Model-free VI	0.05585 ± 0.0007	N/A	N/A

Unsurprisingly, the model-free Dijkstra algorithm has the fastest run time, followed by model-free asynchronous and then synchronous value iteration. The fastest version of Q-learning is purely greedy ($\epsilon = 0$). As a result, we can see that the more randomness we introduce in the form of increasing ϵ, the slower the algorithm runs. However, there is a reverse relationship on the convergence rate: lower ϵ decreases the frequency of convergence. This aligns with intuition because less exploration will prevent states from being visited sufficiently often with high probability. However, as ϵ increases, both the convergence frequency and run time increase.

Interestingly, note that a lower ϵ value causes the optimal cost-to-go from the initial state to converge faster. If our goal is to purely find the optimal path between the initial and the goal and the system is deterministic, picking $\epsilon = 0$ is often the best choice.

The model-free version of Dijkstra's algorithm is 134.65 times faster and uses 22.88 times fewer actions than Q-learning with $\epsilon = 0$. The relationship pattern observed between the different Q-learning under different parameters, and the model-free approaches holds, for all problems we studied, although the difference in overall performance varies.

To properly compare Q-learning to our model-free version of Dijkstra, we use $\epsilon = 0.9$ for which we get consistent convergence of cost-to-go values to their optima over the whole state-space. It is furthermore the fastest parameter

configuration to do so. However, since it is slower than $\epsilon = 0$, the performance difference is increased with model-free Dijkstra which now performs 249.62 times faster and using 46.50 times fewer actions.

3 Analysis of Cost/Reward Models

This section brings the planning-oriented model of Section 2 closer to the models typically used in RL, which involve rewards, discounting, and episodes. We mathematically analyze and experimentally study the effect of each, while remaining in a deterministic setting to improve clarity; Section 4 will then consider the case of stochastic systems.

3.1 Cost and reward models are equivalent, almost

Engineers have tended to minimize *cost* whereas AI researchers, especially in RL, have preferred to maximize *reward*. The former usually involves physical quantities, such as time, energy, distance, or perhaps money in a business setting. The latter takes inspiration from biology by imagining the robot or agent as being conditioned by Pavlovian rewards. As shown below, these formulations are equivalent at a general level; however, the problem with biologically inspired reward is that it tends to lead to the heuristic invention of reward functions to 'motivate' the robot to act in preconceived ways [19]. We therefore introduce the idea of *truecost*, in which the aim is to formulate costs that directly map to the physics of a given system (or perhaps money in a business setting). Ideally, costs or rewards should not be tweaked until the algorithm does what is was supposed to do. Of course, we can imagine *truereward*, but we emphasize that it should be more concrete, such as profit (revenue minus cost) in a business setting, rather than a vague bio-inspired motivator.

It is generally known that replacing the costs in a problem formulation by the corresponding rewards (multiplying the costs by -1) and switching from minimization to maximization results in an equivalent problem formulation in terms of the optimal policies. We show that this is generally true for any cost functional that is linear in terms of the sequence of stepwise costs. Note that this criterion covers all the standard cost formulations, including total cost, discounted cost, asymptotic average cost and the expected values of these in the stochastic formulations.

Proposition 2. *Let $P = (X, U, f, L)$ be a planning problem with a cost functional $L_c : U^{\mathbb{N}} \times X \to \mathbb{R}$. Assume that the cost-functional can be written as $L_c(\tilde{u}, x_1) = \mathcal{L}\big(f(\tilde{u}, x_1)\big)$ in which $f : U^{\mathbb{N}} \times X \to \mathbb{R}^{\mathbb{N}}$ maps a path onto the corresponding sequence of stepwise (immediate) costs and $\mathcal{L} : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ is a linear operator. Then a policy π attains a minimum for L_c if and only if it attains a maximum for the operator L_r defined by $L_r(\tilde{u}, x_1) = \mathcal{L}\big(-f(\tilde{u}, x_1)\big)$.*

Proof. For $x \in X$, let $\tilde{u}^* \in U^{\mathbb{N}}$ be the action sequence corresponding to following an L_c-optimal policy π^* from x. Then $L_c(\tilde{u}^*, x) \leq L_c(\tilde{u}, x)$ for all $\tilde{u} \neq \tilde{u}^*$ so that

$$L_r(\tilde{u}^*, x) = \mathcal{L}\big(-f(\tilde{u}^*, x)\big) \geq \mathcal{L}\big(-f(\tilde{u}, x)\big) = L_r(\tilde{u}, x)$$

for all $\tilde{u} \neq \tilde{u}^*$. This means that applying $\tilde{u}^*$ from x corresponds to maximizing the reward functional L_{r}, in other words, following an L_{r}-optimal policy. The implication in the other direction follows similarly, and since x was arbitrary, the claim follows. □

3.2 The dangers of discounting

Infinite horizon decision making problems tend to yield diverging sums of costs or rewards. If there is a goal, as in the case of planning, then the simplest way to fix it is to use a termination action as introduced in Section 2: No more costs/rewards accumulate after termination, thereby forcing (1) to be finite if the goal is reachable. The preferred way in RL is to use discounted cost/reward, which is a kind of mathematical hack dating back to Bellman [3]. For any $\alpha \in (0, 1)$, the following sum is finite:

$$L_\alpha(\pi, x_1) = \lim_{K \to \infty} \left(\sum_{k=0}^{K} \alpha^k l(x_k, u_k) \right). \tag{8}$$

This implies that future costs/rewards are considered less valuable than current ones. This might be fine in highly unpredictable settings such as financial markets, but for robotics this causes a large and dangerous deviation from truecost.

Let $\tilde{u} = (u_0, \ldots, u_{N-1})$ be a sequence of actions. A sequence of states $(x_0, \ldots, x_N)$ in which $x_{k+1} = f(x_k, u_k)$ for $k = 0, \ldots, N-1$ is a *cycle*, if $x_0 = x_N$ and $x_k \neq x_0$ for $k = 1, \ldots, N-1$. We denote by $C(\tilde{u}, x_0)$ a cycle beginning at x_0 corresponding to $\tilde{u}$. If $\tilde{u}$ is given by a policy π, we denote a cycle starting at x_0 by $C_\pi(x_0)$. We also denote by $\mathcal{C}_P(\pi)$ the collection of cycles generated by a fixed policy π in a planning problem P. For the following proposition, let $\mathcal{P} = (X, U, f, x_1, X_G, \ell)$ be a deterministic planning problem as introduced in Sec. 2. Suppose ℓ is a strictly positive cost function other than $\ell(x, u_T)$ which is 0 for any x. Without loss of generality, X_G is a singleton, that is, there is a single goal state.

Proposition 3. *Given a planning problem $\mathcal{P}$ with an initial state $x_0 = x_I \in X$, suppose there exists a policy π_G for which $L(\pi_G, x_0)$, the cost according to the cost functional given in* (1)*, is finite. Furthermore, suppose there exists a π_C such that the state trajectory $(x_0, \pi_C(x_0), x_1, \pi_C(x_1), \ldots)$ contains a cycle $C_\pi(x_m)$ for some m. Then, there exist ℓ and α such that $L(\pi^*_\alpha, x_1) = \infty$, in which π^*_α is an optimal policy minimizing the cost functional defined in* (8)*.*

(Proof appears in the appendix.)

To illustrate the issue that an optimal solution to a discounted cost problem might result in infinite true cost and miss the goal even if the goal is reachable, consider the planning problem with $X = \{0, 1, \ldots, 5\}$, $x_G = 5$, $U = \{-1, 1\}$, $f(x, u) = \min(5, \max(0, x+u))$, and $\ell(x, u) = x+1$. The only policy that reaches the goal is π_G for which $\pi_G(x) = 1$ for any $x \in X \setminus \{5\}$ and $\pi_G(5) = u_T$. For this policy, the true cost from $x_0 = 0$ is $L(\pi_G, 0) = 15$. Considering the discounted

formulation, let π_C be $\pi_C(x) = -1$ for any $x \in X \setminus \{5\}$ and $\pi_C(5) = u_T$. For this policy, the total discounted cost is $L_\alpha(\pi_C, 0) = 1/(1-\alpha)$ as $K \to \infty$. On the other hand, the total discounted cost associated with the policy that leads to goal, that is, $L_\alpha(\pi_G, 0) = \sum_{i=0}^{4} \alpha^i (i+1)$. Then, if α is selected so that $1/(1-\alpha) < \sum_{i=0}^{4} \alpha^i (i+1)$, an optimal policy π_α^* for the discounted cost formulation will have a cycle. Furthermore, it is easy to see that any optimal policy in this case will have $\pi_\alpha^*(0) = -1$ since any other cycle will incur higher cost. Then, if α is selected accordingly, $L(\pi_\alpha^*, 0) = \lim_{K\to\infty} \sum_{k=0}^{K} 1 = \infty$.

Thus, discounting is a heuristic that is as problematic as using potential functions for motion planning and becoming trapped in local minima [2]. One could try to pick α as close to one as possible, hoping to fend off the problem of ignoring the goal. An alternative way to ensure that the cost remains bounded over an infinite horizon is to optimize the average cost/reward per stage, which is also traced back to Bellman [3], and has been more recently suggested for RL [22,17]. The relationship between this and termination actions is the basis of Section 3.3.

3.3 Episodic equivalences

Imagine that the robot must solve the same tasks over and over, forever. Each time it arrives in the goal, it gets magically teleported back to the initial state and must get to the goal again; this is an example of physical jumping from Section 2. This leads to an infinite-horizon problem that fits the model of Section 2 but is closer to the common RL formulation that relies upon repeated *episodes* [25]. In this section, we establish an equivalence between the two formulations. This is worth considering because the resulting insight extends to problems in which different goal requests may be made in each episode (see, for example, [23]).

Consider the following two problem formulations:

(i) Let $P_{\text{unsp}} = (X, U, f, L, \ell_f)$ be an *unspecified-horizon* problem. We assume there exists a termination action u_T for which $\ell_f(x_g, u_T) = 0$. The cost functional L is given by (1).

(ii) Let $P_{\text{inf}} = (X, U, f, L_{\text{av}})$ be an *infinite-horizon* problem. The cost functional $L_{\text{av}} : U^{\mathbb{N}} \to \mathbb{R}$ is defined as the asymptotic average

$$L_{\text{av}}(\tilde{u}, x_1) := \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \ell(x_k, u_k). \tag{9}$$

Instead of setting $\ell_f(x_g, u_T) = 0$ as in (i), we assume that the system resets to x_I whenever it reaches x_g and incurs a negative cost (a bonus) $M < 0$.

We characterize the planning problems for which the above definitions (i) and (ii) share an optimal policy. For P_{unsp} there exists an optimal state-feedback policy π_{u}^*, which takes the robot from any initial state to the termination state with minimal cost. In P_{inf} there is no termination state, so any optimal policy leads the robot to enter a cycle with the lowest asymptotic average cost and

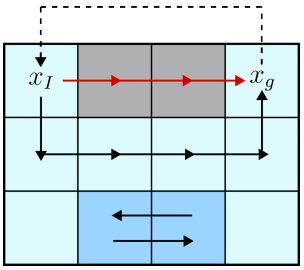

Fig. 3. The difference between a single trial (unspecified horizon) versus infinite horizon formulation. The optimal path for a single trial is indicated with black arrows going from x_I to x_g. A more costly, but shorter, path is indicated with red arrows. In the infinite-horizon formulation, a shortcut (dashed arrow) is introduced that takes the robot from x_g and returns it to x_I. The dark blue squares indicate a region in state space where the robot could loop indefinitely with negligible average cost per step without ever reaching the goal.

to loop there indefinitely. Due to the reset, every cycle containing x_g must also contain x_I. It follows then from the optimality of π_{u}^* that the lowest-cost cycle containing x_g is $C_{\pi_{\mathrm{u}}^*}(x_I)$. However, there may exist other cycles that do not contain x_g and that have a lower asymptotic cost than $L_{\mathrm{av}}(C_{\pi_{\mathrm{u}}^*}(x_I))$. In such cases π_{u}^* fails to be optimal for P_{inf}. This can be remedied by tuning the bonus parameter M so that its cumulative effect overcomes the difference. However, changing the magnitude of M may also have the effect that some cycle $C(\tilde{u}, x_I) \neq C_{\pi_{\mathrm{u}}^*}(x_I)$ containing x_g may become optimal under the asymptotic average-cost formulation due to the resetting. See Figure 3 for an illustration of the two types of cycles. The conclusion is that the formulations P_{unsp} and P_{inf} are not generally equivalent, but may be so if certain conditions regarding the lengths and average costs of cycles in P_{inf} are satisfied. This is made precise in the following.

For an action sequence $\tilde{u}_k := (u_0, \ldots, u_{k-1})$, denote by $\Phi\big(x, \tilde{u}_k\big)$ the path $(x_0, \ldots, x_k)$ in which $x_0 = x$ and $x_{j+1} = f(x_j, u_j)$ for all $j = 0, \ldots, k-1$. Denote by $\varphi(x, \tilde{u}_k)$ the last state in $\Phi(x, \tilde{u}_k)$. Let

$$U_G := \{\tilde{u} : \varphi(x_I, \tilde{u}) = x_g\}$$

be the set of action sequences that take the robot from x_I to x_g. The policies for P_{inf} can be divided into those that correspond to applying some $\tilde{u} \in U_G$ from x_I, and those that tend towards some other cycle that does not contain x_g. For any policy π that corresponds to looping an action sequence $\tilde{u} \in U_G$ from the initial state x_I, the asymptotic average cost is given by

$$L_{\mathrm{av}}(\tilde{u}, x_I) = \frac{M + L(\tilde{u}, x_I)}{S(\tilde{u})} = \frac{1}{S(\tilde{u})} \left(M + \sum_{k=0}^{S(\tilde{u})-1} \ell\big(x_k, f(x_k, u_k)\big) \right) \tag{10}$$

in which $S(\tilde{u})$ denotes the length of the path $\Phi\big(x_I, \tilde{u}\big)$. Let $\tilde{u}^* \in U_G$ be the action sequence corresponding to applying the optimal policy for P_{unsp} at x_I. Define

for each $\tilde{u} \in U_G$ the value

$$\mathcal{A}(\tilde{u}) := \left(1 - \frac{S(\tilde{u})}{S(\tilde{u}^*)}\right)\left(\frac{S(\tilde{u})L(\tilde{u}^*, x_I)}{S(\tilde{u}^*)L(\tilde{u}, x_I)} - 1\right) \tag{11}$$

and the sets

$$\mathcal{R}^- := \big\{\tilde{u} : S(\tilde{u}) < S(\tilde{u}^*)\big\}, \qquad \mathcal{R}^+ := \big\{\tilde{u} : S(\tilde{u}) > S(\tilde{u})^*\big\}. \tag{12}$$

Furthermore, let $\hat{C} = C(\tilde{u}, x_0)$ be the cycle which has, among all the cycles that do not contain x_g, the smallest asymptotic average cost

$$L_{\text{av}}(\hat{C}) = L_{\text{av}}(\tilde{u}, x_0) = \frac{L(\tilde{u}, x_0)}{S(\tilde{u}) - 1}. \tag{13}$$

Proposition 4. *Let π^* be an optimal policy for P_{unsp}. Then π^* is also an optimal policy for P_{inf} with reset bonus $M < 0$ if and only if M satisfies*

$$\alpha \leq M \leq \min\big\{\beta, S(\tilde{u}^*)L_{av}(\hat{C}) - L(\tilde{u}^*, x_I)\big\},$$

in which $\tilde{u}^*$ is the action sequence corresponding to applying π^* from x_I to reach x_g, $\hat{C}$ is as in (13)*, $\mathcal{A}(\tilde{u})$ is as in* (11)*, and*

$$\alpha := \max_{\tilde{u} \in \mathcal{R}^-}\{\mathcal{A}(\tilde{u})\}, \quad \textit{and} \quad \beta := \min_{\tilde{u} \in \mathcal{R}^+}\{\mathcal{A}(\tilde{u})\}.$$

(Proof appears in the appendix.)

4 From Deterministic to Stochastic Models

4.1 Basic assumptions and approach

The concepts from Sections 2 and 3 extend gracefully to the stochastic setting by replacing f in $\mathcal{P}$ with a probabilistic transition model $P(x_{k+1}|x_k, u_k)$, which denotes the probability that x_{k+1} is reached if u_k is applied from x_k. This causes paths taken during execution to be unpredictable, resulting in G^* and Q^* representing *expected* costs. Furthermore, every state-action pair should be visited an infinite number of times to learn P, rather than once to learn f from the deterministic setting. Thus, the method from Section 2 of quickly learning f first, and then applying Dijkstra's algorithm is no longer applicable. At best, one could try to estimate the transition probabilities to some level of statistical confidence and then apply stochastic value iteration to obtain G^* and π^*. For comparative purposes, we will consider the case of stochastic value iteration applied to a perfectly learned transition model. However, most of our studies in this section are based on Q-learning.

In our implementations, the transition model $P(x_{k+1}|x_k, u_k)$ is defined to be a direct extension of f from Section 2, in terms of a *predictability factor*, $\gamma \in (0, 1]$, that controls the amount of entropy or prediction uncertainty. From

each x_k, consider applying u_k to get $x_{k+1} = f(x_k, u_k)$. For the stochastic version with predictability factor γ, we define $P(x'_{k+1}|x_k, u_k) = \gamma$ if $x'_{k+1} = x_{k+1}$; otherwise, the remaining probability mass $1 - \gamma$ is spread uniformly as if nature causes the remaining actions (excluding u_k) to be chosen from $U(x_k)$ (including a termination action that holds the state constant). If $\gamma = 1$, then the deterministic model can essentially be simulated. If $\gamma = 1/2$, then on average, half of the time the robot moves in the direction it was commanded, and for the other half it effectively executes an alternative action randomly.

Another critical parameter is the learning rate ρ, which appears in (5). Imagine implementing a complementary filter that uses ρ to interpolate between a noisy data source and the current best estimate. If the data is perfectly reliable, as in the deterministic case of Section 2, then we can set $\rho = 1$. As the predictability decreases (lower γ), we would expect to use lower values for ρ so that the noise is attenuated and the estimate becomes more stable. Choosing a good value of ρ is a delicate art. We considered multiple values in our experiments, and furthermore developed an adaptive tuning method for ρ so that it decreases during execution according to the formula

$$\rho = \frac{1}{n(x, u)^{\omega}}, \tag{14}$$

in which $n(x, u)$ is the number of times that each state-action pair has been visited and ω is an arbitrary parameter (this idea was introduced in [9] and used in the Double Q-learning approach [10]).

4.2 Computational comparisons

The experimental setup outlined in Section 2.3 is reused with a few minor changes. In cases where $\gamma \leq 0.9$ we use 10 samples instead of 100. We increased the number of episodes to 3000 to improve the convergence frequency, that is, the percentage of times the values converge. We investigated performance by trading off three parameters: learning rate ρ, greedy parameter ϵ, and predictability factor γ. We studied convergence rates in comparison to the true optimal cost to go, obtained from stochastic value iteration, and flagged cases in which the Q-values are within 10% of optimal to identify approximate successes.

For cases of high γ (more predictable), we investigated combinations of $\epsilon, \rho \in \{0.5, 0.7, 0.9, 0.99\}$. The ρ values were chosen after careful investigation and observing that for $\rho < 0.5$ and $\rho > 0.99$, the results were not more informative. We also found that increasing ρ decreases run time. In Table 2, which combines the data for Problems 1 and 10, note that for $\gamma = 0.999$ there is less benefit of being greedy in comparison to the results of Section 2.3. Although choosing $\epsilon = 0$ still leads to a better or equal run time to $\epsilon = 1$, we see that $\epsilon = 0.9$ has a much better performance than both, while also converging. This appears to be due the fact that nondeterminism throws the state away from greedy routes, thereby requiring more exploration to get values to converge. Furthermore, we can see that dynamic programming methods converge about

two orders of magnitude more quickly than RL, which reflects the price of learning on the fly. Similar trends were observed in other problems; the appendix contains many more. It also reports statistics for the goal discovery time, which are similar to the deterministic case, but grow with increasing ρ.

Table 2. Performance of Q-learning and dynamic programming methods for Problems 1 and 10 for stochastic problem with $\gamma = 0.999$. Note the addition of two columns: how often the value for the initial state converges; the shortest and longest path achieved over all samples.

Algorithm ($\gamma = 0.999$)	Run time (mean ± std)	Convergence	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
(Pr 1) Q-learning ($\rho = 0.99, \epsilon = 0$)	0.40677 ± 0.0041	0.0%	0.0888 ± 0.0008	100.0%	28/32
(Pr 1) Q-learning ($\rho = 0.99, \epsilon = 0.25$)	0.52970 ± 0.0064	0.0%	0.1006 ± 0.0024	100.0%	28/29
(Pr 1) Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.76335 ± 0.0084	0.0%	0.1144 ± 0.0024	100.0%	28/30
(Pr 1) Q-learning ($\rho = 0.99, \epsilon = 0.75$)	0.52489 ± 0.1930	100.0%	0.1516 ± 0.0045	100.0%	28/30
(Pr 1) Q-learning ($\rho = 0.99, \epsilon = 0.9$)	0.27856 ± 0.0271	100.0%	0.2044 ± 0.0109	100.0%	28/29
(Pr 1) Q-learning ($\rho = 0.99, \epsilon = 1$)	0.40572 ± 0.0491	100.0%	0.3512 ± 0.0450	100.0%	28/30
(Pr 1) Async Value Iteration	0.12869 ± 0.0009	N/A	N/A	N/A	28/29
(Pr 1) Value Iteration	0.17870 ± 0.0024	N/A	N/A	N/A	28/30
(Pr 10) Q-learning ($\rho = 0.99, \epsilon = 0$)	0.94648 ± 0.0174	0.0%	0.2290 ± 0.0314	100.0%	64/27731
(Pr 10) Q-learning ($\rho = 0.99, \epsilon = 0.25$)	1.24982 ± 0.0203	0.0%	0.2513 ± 0.0056	100.0%	64/66
(Pr 10) Q-learning ($\rho = 0.99, \epsilon = 0.5$)	1.91614 ± 0.0188	0.0%	0.3029 ± 0.0058	100.0%	64/67
(Pr 10) Q-learning ($\rho = 0.99, \epsilon = 0.75$)	0.97765 ± 0.2121	100.0%	0.4122 ± 0.0107	100.0%	64/66
(Pr 10) Q-learning ($\rho = 0.99, \epsilon = 0.9$)	0.65742 ± 0.0572	100.0%	0.5611 ± 0.0225	100.0%	64/66
(Pr 10) Q-learning ($\rho = 0.99, \epsilon = 1$)	2.26297 ± 0.7142	100.0%	1.6362 ± 0.4014	100.0%	64/66
(Pr 10) Async Value Iteration	0.21926 ± 0.0162	N/A	N/A	N/A	64/66
(Pr 10) Value Iteration	0.35463 ± 0.0252	N/A	N/A	N/A	64/66

By lowering γ to 0.99, we observe that convergence no longer happens, even if we increase the episode count to 5000 or 10000. In fact, upon further investigation, the average distance between the optimal values and the values obtained with Q-learning stayed nearly constant while changing episodes when $\epsilon = 1$, which means that there is a convergence happening on a state-space level, but to different values than the optimal ones. However, the initial cost-to-go values still converge for some configurations. Table 3 shows when this happens for Problems 1 and 10. The general observation is that a lower ρ yields more reliable convergence because it is more robust to unpredictability, but requires more iterations. As γ is lowered, then ρ should be lowered accordingly, although the best practice is found only by trial and error. It is interesting to note that a policy that leads the robot to the goal in a nearly optimal manner can be computed even if the Q-values from the initial state do not converge, as seen in the configuration with $\rho = \epsilon = 0.5$ for Problem 1.

Interestingly, as we decrease γ, we observed for Problems 1 and 11 that the run time of value iteration becomes closer to that of greedy Q-learning with a low ρ.

To study global state-space convergence to optimal values, we investigated the effects of a low ρ on simpler problems (i.e., Problem 8 and Problem 16). For $\gamma = 0.9$ we found that a random policy ($\epsilon = 1$) with $\rho = 0.1$ converged for both problems, whereas for Problem 16 even a greedy policy ($\epsilon = 0$) with the same ρ

Table 3. Performance of Q-learning and dynamic programming methods for Problems 1 and 10 for stochastic problem with $\gamma = 0.99$.

Algorithm ($\gamma = 0.99$)	Run time (mean ± std)	Convergence	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
(Pr 1) Q-learning ($\rho = 0.5, \epsilon = 0$)	0.44775 ± 0.0094	0.0%	0.1815 ± 0.0082	100.0%	28/34
(Pr 1) Q-learning ($\rho = 0.99, \epsilon = 0$)	0.42034 ± 0.0090	0.0%	0.2078 ± 0.0750	86.0%	28/2205
(Pr 1) Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.82059 ± 0.0169	0.0%	0.5143 ± 0.1522	32.0%	28/31
(Pr 1) Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.77585 ± 0.0109	0.0%	nan ± nan	0.0%	28/1474
(Pr 1) Q-learning ($\rho = 0.5, \epsilon = 1$)	15.46274 ± 0.1914	0.0%	5.2842 ± 3.5571	90.0%	28/33
(Pr 1) Q-learning ($\rho = 0.99, \epsilon = 1$)	15.39603 ± 0.1846	0.0%	nan ± nan	0.0%	28/32
(Pr 1) Async Value Iteration	0.15656 ± 0.0014	N/A	N/A	N/A	28/33
(Pr 1) Value Iteration	0.22934 ± 0.0020	N/A	N/A	N/A	28/31
(Pr 10) Q-learning ($\rho = 0.5, \epsilon = 0$)	1.09601 ± 0.0226	0.0%	0.4524 ± 0.0133	100.0%	64/322
(Pr 10) Q-learning ($\rho = 0.99, \epsilon = 0$)	1.07131 ± 0.0178	0.0%	0.2651 ± 0.0000	1.0%	66/2596
(Pr 10) Q-learning ($\rho = 0.5, \epsilon = 0.5$)	2.08346 ± 0.0157	0.0%	0.7959 ± 0.2102	100.0%	64/68
(Pr 10) Q-learning ($\rho = 0.99, \epsilon = 0.5$)	1.99043 ± 0.0217	0.0%	0.6288 ± 0.3471	99.0%	64/3689
(Pr 10) Q-learning ($\rho = 0.5, \epsilon = 1$)	30.93903 ± 0.2059	0.0%	3.5574 ± 1.1094	100.0%	64/70
(Pr 10) Q-learning ($\rho = 0.99, \epsilon = 1$)	30.94423 ± 0.1896	0.0%	2.4456 ± 1.0780	100.0%	64/610
(Pr 10) Async Value Iteration	0.24640 ± 0.0036	N/A	N/A	N/A	64/69
(Pr 10) Value Iteration	0.40702 ± 0.0028	N/A	N/A	N/A	64/70

can suffice. However, this is approach is not enough for $\gamma = 0.7$. Instead, we were only able to obtain global convergence with $\epsilon = 1$ by decaying ρ as described in (14), using $\omega = 0.7$. We obtained global convergence for Problem 16 with $\gamma = 0.5$ using the same approach. We were unable to get global convergence in Problem 8 by any choice of ω.

5 Conclusions

Our work has brought planning and reinforcement learning into closer alignment so that approaches, assumptions, and models across both can be compared and be better understood. Through the introduction of a fully deterministic version of RL, we established its convergence and studied its performance relative to dynamic-programming based planning. Through analysis of cost models, we have warned against using discounted cost, as is popular in RL; instead, we have advocated for truecost, in which the costs/rewards translate directly into physical or monetary values, rather than using them as a way to tune performance. We have also established equivalences of maximizing rewards and minimizing costs, and episodic models to single-shot goal satisfaction models. These provide motivation to use undiscounted cost models and termination actions for RL for solving goal-oriented tasks. We applied these models in the stochastic system setting and showed how they extend naturally from the deterministic case, but inherit unique RL-oriented challenges, such as greediness and learning rate parameter tuning, which warrant further study. We are currently extending the mathematical analysis of cost-reward and episodic equivalences to the stochastic case.

Acknowledgments We thank Markku Suomalainen for helpful feedback and discussions.

References

1. S. Albers and M. T. Henzinger. Exploring unknown environments. *SIAM Journal on Computing*, 29(4):1164–1188, 2000.
2. J. Barraquand, B. Langlois, and J. C. Latombe. Numerical potential field techniques for robot path planning. *IEEE Transactions on Systems, Man, & Cybernetics*, 22(2):224–241, 1992.
3. R. E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.
4. D. P. Bertsekas. Distributed asynchronous computation of fixed points. *Mathematical Programming*, 27(1):107–120, 1983.
5. D. P. Bertsekas. *Reinforcement learning and optimal control*, volume 1. Athena Scientific, 2019.
6. D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, 1996.
7. H.-T. L. Chiang, J. Hsu, M. Fiser, L. Tapia, and A. Faust. Rl-rrt: Kinodynamic motion planning via learning reachability estimators from rl policies. *IEEE Robotics and Automation Letters*, 4:4298–4305, 2019.
8. G. Dudek, M. Jenkin, E. Milios, and D. Wilkes. Map validation and self-location in a graph-like world. In *Proceedings AAAI National Conference on Artificial Intelligence*, pages 1648–1653, 1993.
9. E. Even-Dar and Y. Mansour. *Learning Rates for Q-Learning*, page 589–604. Springer Berlin Heidelberg, 2001.
10. H. Hasselt. Double q-learning. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.
11. R. Kalman. Randomness reexamined. *Modeling, Identification and Control*, 15(3):141–151, 1994.
12. J. Kober, J. A. Bagnell, and J. Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.
13. S. Koenig and M. Likhachev. D^* lite. In *Proceedings AAAI National Conference on Artificial Intelligence*, pages 476–483, 2002.
14. S. M. LaValle. *Planning Algorithms*. Cambridge University Press, Cambridge, U.K., 2006. Also available at http://lavalle.pl/planning/.
15. S. M. LaValle, M. S. Branicky, and S. R. Lindemann. On the relationship between classical grid search and probabilistic roadmaps. *International Journal of Robotics Research*, 23(7/8):673–692, July/August 2004.
16. S. M. LaValle and P. Konkimalla. Algorithms for computing numerical optimal feedback motion strategies. *International Journal of Robotics Research*, 20(9):729–752, September 2001.
17. S. Mahadevan. Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine Learning*, 22(1–3):159–195, 1996.
18. A. Naik, R. Shariff, N. Yasui, H. Yao, and R. S. Sutton. Discounted reinforcement learning is not an optimization problem, 2019. arXiv, 1910.02140.
19. A. Y. Ng, D. Harada, and S. J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, page 278–287, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.
20. M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, April 1994.

21. H. Robbins and S Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, September 1951.
22. A. Schwartz. *A Reinforcement Learning Method for Maximizing Undiscounted Rewards*, page 298–305. Elsevier, 1993.
23. R. Sharma, S. M. LaValle, and S. A. Hutchinson. Optimizing robot motion strategies for assembly with stochastic models of the assembly process. *IEEE Trans. on Robotics and Automation*, 12(2):160–174, April 1996.
24. A. Stentz. Optimal and efficient path planning for partially-known environments. In *Proceedings IEEE International Conference on Robotics & Automation*, pages 3310–3317, 1994.
25. R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.
26. C. Tang, B. Abbatematteo, J. Hu, R. Chandra, R. Martín-Martín, and P. Stone. Deep reinforcement learning for robotics: A survey of real-world successes. *Annual Review Control, Robotics, and Autonomous Systems*, 8:153–188, 2025.
27. K. G. Timperi, A. J. LaValle, and S. M. LaValle. Universal plans: One action sequence to solve them all! In N. Amato, K. Driggs-Campbell, E. Chinwe, M. Morales, and J. M. O'Kane, editors, *Algorithmic Foundations of Robotics, XVI*. Springer-Verlag, Berlin, 2025.
28. C. J. Watkins and P Dayan. Technical note: Q-learning. *Machine Learning*, 8(3–4):279–292, May 1992.
29. D. Yershov and S. M. LaValle. Simplicial Dijkstra and A* algorithms for optimal feedback planning. In *Proceedings IEEE International Conference on Intelligent Robots and Systems*, 2011.

A Proofs of Propositions

Proof of Proposition 3

Proof. Without loss of generality suppose that the cycle given by π_C starts at x_I, that is $C_{\pi_C}(x_I) = (x_0 = x_I, x_1, \ldots, x_N = x_I)$ for some N. Furthermore, suppose the cycle does not include x_G. Note that since π_C is a policy and we are considering an infinite-horizon setting, this cycle will be executed infinitely many times. Let $\bar{c} = c_0 + \alpha c_1 + \alpha^2 c_2 + \ldots \alpha^{N-1} c_{N-1}$ be the cost that accumulates going through this cycle once, in which $c_i = \ell(x_i, \pi_C(x_i))$. Let $\bar{\alpha} := \alpha^N$. The discounted cost associated with π_C is

$$L_\alpha(\pi_C, x_I) = \lim_{K \to \infty} \sum_{i=0}^{K} \bar{\alpha}^i \bar{c} = \frac{\bar{c}}{1 - \bar{\alpha}}.$$

Let π^*_G be a policy that terminates at goal and minimizes the discounted cost functional in (8) that is,

$$\pi^*_G := \arg\min_{\pi \in \Pi} \{L_\alpha(\pi, x_I) \,|\, \exists K < |X| \wedge x_K = x_G\},$$

in which, Π is the set of all policies and x_K is the state reached from $x_0 = x_I$ by determining $x_{i+1} = f(x_i, \pi(x_i))$ for $i = 0, \ldots, K-1$. The total discounted cost for this policy is $L_\alpha(\pi^*_G, x_I) = \sum_{k=0}^{K-1} \alpha^k \ell(x_k, \pi^*_G(x_k))$. Notice that this is a finite sum since the cost of termination action u_T, which is applied at the goal state for $k > K-1$, is 0. Furthermore, this is a lower bound on the policies that reach the goal state, that is, that have finite true cost. Denote by π^*_α, an optimal policy minimizing the cost functional given in (8). Then, $L_\alpha(\pi^*_\alpha, x_I) \leq L_\alpha(\pi_C, x_I)$ and $L_\alpha(\pi^*_\alpha, x_I) \leq L_\alpha(\pi^*_G, x_I)$. Because $L_\alpha(\pi^*_\alpha, x_I)$ is upper bounded, if $L_\alpha(\pi^*_\alpha, x_I) = \frac{\bar{c}}{1-\bar{\alpha}} < \sum_{k=0}^{K-1} \alpha^k \ell(x_k, \pi^*_G(x_k)) = L_\alpha(\pi^*_G, x_I)$, π^*_α must contain a cycle. Then, we can pick ℓ and α such that this inequality is satisfied. In that case, $L(\pi^*_\alpha, x_I) = \infty$ since it contains a cycle. □

Proof of Proposition 4

Proof. For π^* to be optimal for P_{inf}, the asymptotic average cost $L_{\text{av}}(\tilde{u}^*, x_I) = \big(M + L(\tilde{u}^*, x_I)\big)/S(\tilde{u}^*)$ must attain the lowest value among both the action sequences $\tilde{u} \in U_G$ and $\tilde{u} \notin U_G$. The second condition holds if and only if the cost $L_{\text{av}}(\tilde{u}^*, x_I)$ is less than $L_{\text{av}}(\hat{C})$, which can be expressed as

$$M < S(\tilde{u}) L_{\text{av}}(\hat{C}) - L(\tilde{u}, x_I).$$

For $\tilde{u}^*$ to yield the lowest asymptotic average cost among the action sequences $\tilde{u} \in U_G$, it needs to satisfy the inequality

$$\frac{L(\tilde{u}^*, x_I) + M}{S(\tilde{u}^*)} \leq \frac{L(\tilde{u}, x_I) + M}{S(\tilde{u})} \tag{15}$$

for every $\tilde{u} \in U_G$. This is equivalent to $M \leq \mathcal{A}(\tilde{u})$ for $\tilde{u} \in \mathcal{R}^+$ and to $M \geq \mathcal{A}(\tilde{u})$ for $\tilde{u} \in \mathcal{R}^-$. Thus, for $\tilde{u}^*$ to satisfy (15) relative to all $\tilde{u} \in U_G$ is equivalent to M satisfying the double inequality $\max_{\tilde{u} \in \mathcal{R}^-} \{\mathcal{A}(\tilde{u})\} \leq M \leq \min_{\tilde{u} \in \mathcal{R}^+} \{\mathcal{A}(\tilde{u})\}$. □

B Problems

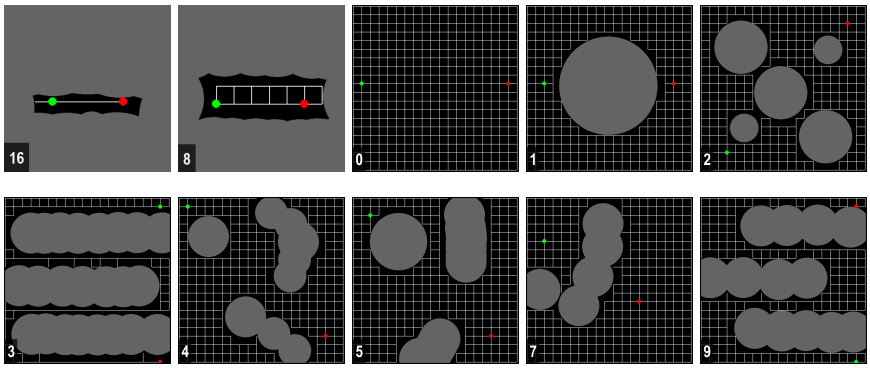

Fig. 4. Ten problems (enumerated from top-left to bottom-right as 16, 8, 0, 1, 2, 3, 4, 5, 7, 9). The green dot is the initial state, while the red dot is the goal state. The white lines connect the states and can be viewed as the actions allowed by the robot. The gray circles represent obstacles that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

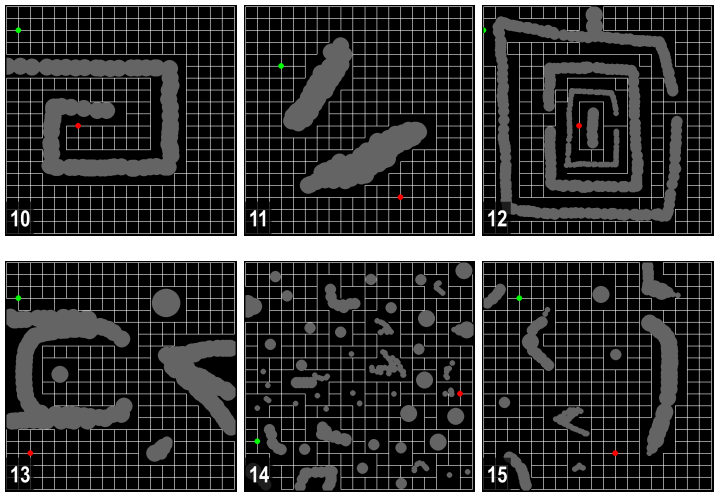

Fig. 5. Additional six problems (enumerated from top-left to bottom-right as 10 - 15).

C 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 ϵ-decay.

Algorithm (Problem 0)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)
Q-learning ($\epsilon = 0$)	0.11326 ± 0.0007	32774.0000 ± 0.0000	0.0 %	0.0075 ± 0.0000	2131.0000 ± 0.0000	0.0622 ± 0.0003	18000.0000 ± 0.0000
Q-learning ($\epsilon = 0.25$)	0.15533 ± 0.0019	45186.3700 ± 270.0393	0.0 %	0.0138 ± 0.0019	4041.3800 ± 537.1843	0.0691 ± 0.0014	20140.0000 ± 374.6999
Q-learning ($\epsilon = 0.5$)	0.24147 ± 0.0025	72453.8400 ± 618.0135	0.0 %	0.0181 ± 0.0029	5436.8200 ± 871.1261	0.0831 ± 0.0023	24990.0000 ± 655.6676
Q-learning ($\epsilon = 0.75$)	0.58168 ± 0.0079	180569.8300 ± 2367.9754	0.0 %	0.0199 ± 0.0048	6167.2200 ± 1481.6033	0.1117 ± 0.0038	34790.0000 ± 1210.7436
Q-learning ($\epsilon = 0.9$)	0.95443 ± 0.1644	303767.8100 ± 52175.4497	100.0 %	0.0158 ± 0.0081	5039.6800 ± 2570.6394	0.1504 ± 0.0076	48030.0000 ± 2451.3466
Q-learning ($\epsilon = 1$)	0.40398 ± 0.0332	135913.8400 ± 11113.1169	100.0 %	0.0050 ± 0.0039	1636.8800 ± 1303.7126	0.2082 ± 0.0192	69980.0000 ± 6452.8753
Q-learning (π) ($\epsilon = 1$)	0.41524 ± 0.0313	136706.6400 ± 10296.1120	100.0 %	0.0042 ± 0.0034	1359.0800 ± 1101.2584	0.2114 ± 0.0187	69530.0000 ± 6119.5670
Model-free Dijkstra	0.00295 ± 0.0005	5368.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Model-free Asynchronous Value Iteration	0.02134 ± 0.0005	31654.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Model-free Value Iteration	0.03276 ± 0.0007	48374.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan

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

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

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

Algorithm (Problem 4)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)
Q-learning ($\epsilon = 0$)	0.15125 ± 0.0034	46110.0000 ± 0.0000	0.0 %	0.0044 ± 0.0003	1325.0000 ± 0.0000	0.0724 ± 0.0021	22000.0000 ± 0.0000
Q-learning ($\epsilon = 0.25$)	0.19840 ± 0.0022	60374.1400 ± 315.4138	0.0 %	0.0147 ± 0.0037	4472.6600 ± 1134.9649	0.0869 ± 0.0021	26480.0000 ± 538.1450
Q-learning ($\epsilon = 0.5$)	0.28498 ± 0.0053	88913.2900 ± 649.2647	0.0 %	0.0177 ± 0.0043	5566.3400 ± 1348.1139	0.1081 ± 0.0039	33890.0000 ± 1038.2196
Q-learning ($\epsilon = 0.75$)	0.57889 ± 0.0082	186567.6800 ± 2407.6892	0.0 %	0.0203 ± 0.0061	6607.7800 ± 1982.5550	0.1520 ± 0.0090	49330.0000 ± 2956.5351
Q-learning ($\epsilon = 0.9$)	1.36798 ± 0.2875	451724.6400 ± 94716.2085	58.0 %	0.0168 ± 0.0098	5610.9200 ± 3275.1360	0.2096 ± 0.0166	69800.0000 ± 5499.0908
Q-learning ($\epsilon = 1$)	0.80289 ± 0.1904	278558.3600 ± 65909.6114	100.0 %	0.0085 ± 0.0086	2930.4200 ± 2986.4306	0.3581 ± 0.0701	124770.0000 ± 24373.2866
Q-learning (π) ($\epsilon = 1$)	0.81879 ± 0.2648	277054.3300 ± 88885.4437	100.0 %	0.0076 ± 0.0073	2578.1200 ± 2461.4350	0.3647 ± 0.0628	124020.0000 ± 21413.5378
Model-free Dijkstra	0.00207 ± 0.0005	3663.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Model-free							
Asynchronous Value Iteration	0.01690 ± 0.0005	24447.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Model-free Value Iteration	0.02575 ± 0.0005	37837.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan

Algorithm (Problem 5)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)
Q-learning ($\epsilon = 0$)	0.14089 ± 0.0011	42386.0000 ± 0.0000	0.0 %	0.0047 ± 0.0001	1430.0000 ± 0.0000	0.0625 ± 0.0004	19000.0000 ± 0.0000
Q-learning ($\epsilon = 0.25$)	0.18808 ± 0.0012	56130.6800 ± 264.3807	0.0 %	0.0107 ± 0.0029	3236.7400 ± 879.1809	0.0746 ± 0.0019	22440.0000 ± 571.3143
Q-learning ($\epsilon = 0.5$)	0.27472 ± 0.0027	84451.7000 ± 702.8104	0.0 %	0.0139 ± 0.0029	4341.1200 ± 894.5730	0.0940 ± 0.0044	29150.0000 ± 1336.9742
Q-learning ($\epsilon = 0.75$)	0.58191 ± 0.0116	184877.3400 ± 2306.1895	0.0 %	0.0152 ± 0.0044	4888.6000 ± 1401.3910	0.1350 ± 0.0120	43400.0000 ± 3778.8887
Q-learning ($\epsilon = 0.9$)	0.98171 ± 0.2661	319656.4200 ± 86716.4391	98.0 %	0.0112 ± 0.0065	3673.4600 ± 2140.2486	0.1834 ± 0.0206	60150.0000 ± 6788.7775
Q-learning ($\epsilon = 1$)	0.49500 ± 0.1111	170925.8200 ± 38202.6919	100.0 %	0.0058 ± 0.0048	1977.1400 ± 1657.1752	0.2829 ± 0.0489	97870.0000 ± 16768.8133
Q-learning (π) ($\epsilon = 1$)	0.53800 ± 0.1353	179624.1400 ± 44847.8988	100.0 %	0.0060 ± 0.0046	2018.1600 ± 1544.2775	0.2987 ± 0.0459	100000.0000 ± 15316.0047
Model-free Dijkstra	0.00204 ± 0.0002	3615.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Model-free							
Asynchronous Value Iteration	0.01620 ± 0.0007	23059.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Model-free Value Iteration	0.02471 ± 0.0006	36241.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan

Algorithm (Problem 7)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)
Q-learning ($\epsilon = 0$)	0.12079 ± 0.0007	37478.0000 ± 0.0000	0.0 %	0.0029 ± 0.0000	880.0000 ± 0.0000	0.0459 ± 0.0004	14000.0000 ± 0.0000
Q-learning ($\epsilon = 0.25$)	0.16239 ± 0.0016	49680.7600 ± 323.7531	0.0 %	0.0050 ± 0.0018	1508.2000 ± 544.4474	0.0538 ± 0.0015	16270.0000 ± 443.9595
Q-learning ($\epsilon = 0.5$)	0.24275 ± 0.0027	76035.4800 ± 661.8601	0.0 %	0.0068 ± 0.0023	2071.0600 ± 695.9548	0.0669 ± 0.0022	20780.0000 ± 672.0119
Q-learning ($\epsilon = 0.75$)	0.52271 ± 0.0068	169471.7200 ± 2123.6347	0.0 %	0.0087 ± 0.0036	2803.9400 ± 1157.2607	0.0927 ± 0.0042	29970.0000 ± 1374.4453
Q-learning ($\epsilon = 0.9$)	1.50883 ± 0.1231	499216.7200 ± 40690.9169	34.0 %	0.0077 ± 0.0047	2541.0600 ± 1558.5223	0.1312 ± 0.0111	43450.0000 ± 3691.5444
Q-learning ($\epsilon = 1$)	1.82401 ± 0.3753	633100.6200 ± 130431.7955	100.0 %	0.0081 ± 0.0078	2820.5200 ± 2756.1423	0.3153 ± 0.0780	109720.0000 ± 27010.0278
Q-learning (π) ($\epsilon = 1$)	1.96900 ± 0.6300	662812.2800 ± 211950.3269	100.0 %	0.0095 ± 0.0107	3198.5800 ± 3640.9636	0.3005 ± 0.0670	101600.0000 ± 22493.5546
Model-free Dijkstra	0.00217 ± 0.0005	3856.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Model-free							
Asynchronous Value Iteration	0.01852 ± 0.0005	26437.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Model-free Value Iteration	0.02888 ± 0.0001	42649.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan

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

Algorithm (Problem 9)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)
Q-learning ($\epsilon = 0$)	0.30094 ± 0.0025	93512.0000 ± 0.0000	0.0 %	0.0030 ± 0.0000	924.0000 ± 0.0000	0.1071 ± 0.0010	33000.0000 ± 0.0000
Q-learning ($\epsilon = 0.25$)	0.40332 ± 0.0026	124056.6000 ± 369.5486	0.0 %	0.0063 ± 0.0049	1920.4400 ± 1473.8903	0.1266 ± 0.0022	38730.0000 ± 580.6031
Q-learning ($\epsilon = 0.5$)	0.59771 ± 0.0045	189069.5400 ± 805.8806	0.0 %	0.0103 ± 0.0063	3219.8200 ± 1985.7043	0.1528 ± 0.0037	48240.0000 ± 1132.4310
Q-learning ($\epsilon = 0.75$)	0.32806 ± 0.0497	106893.5400 ± 16033.8762	100.0 %	0.0160 ± 0.0085	5189.9200 ± 2773.5873	0.2082 ± 0.0080	67830.0000 ± 2561.4644
Q-learning ($\epsilon = 0.9$)	0.30673 ± 0.0195	102077.7200 ± 6343.1919	100.0 %	0.0330 ± 0.0177	10934.4600 ± 5863.3670	0.2969 ± 0.0170	98810.0000 ± 5556.4287
Q-learning ($\epsilon = 1$)	1.74662 ± 0.7671	611856.5000 ± 269104.1334	100.0 %	0.1213 ± 0.1074	42628.7200 ± 37559.9343	1.2952 ± 0.4267	453590.0000 ± 149158.4456
Q-learning (π) ($\epsilon = 1$)	2.10442 ± 1.1306	704462.5600 ± 377424.9032	100.0 %	0.1779 ± 0.1123	59863.2600 ± 37681.3580	1.3004 ± 0.2590	435550.0000 ± 87139.9306
Model-free Dijkstra	0.00143 ± 0.0005	2360.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Model-free							
Asynchronous Value Iteration	0.02774 ± 0.0002	38776.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Model-free Value Iteration	0.03483 ± 0.0002	50692.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan

Algorithm (Problem 10)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)
Q-learning ($\epsilon = 0$)	0.33383 ± 0.0025	101214.0000 ± 0.0000	0.0 %	0.0068 ± 0.0002	2015.0000 ± 0.0000	0.1720 ± 0.0016	52000.0000 ± 0.0000
Q-learning ($\epsilon = 0.25$)	0.44088 ± 0.0030	133171.4600 ± 405.0955	0.0 %	0.0101 ± 0.0058	3050.0200 ± 1747.4311	0.2043 ± 0.0041	61830.0000 ± 1200.4582
Q-learning ($\epsilon = 0.5$)	0.64394 ± 0.0045	200028.6600 ± 996.3678	0.0 %	0.0168 ± 0.0071	5222.4200 ± 2184.3765	0.2517 ± 0.0066	78540.0000 ± 1951.5122
Q-learning ($\epsilon = 0.75$)	1.08726 ± 0.2099	348322.6400 ± 66374.2689	76.0 %	0.0231 ± 0.0100	7449.8000 ± 3206.2196	0.3502 ± 0.0129	112990.0000 ± 3410.2639
Q-learning ($\epsilon = 0.9$)	0.61889 ± 0.0559	205682.0500 ± 18488.2891	100.0 %	0.0372 ± 0.0194	12384.0200 ± 6398.5418	0.4858 ± 0.0239	161590.0000 ± 7636.8776
Q-learning ($\epsilon = 1$)	1.88356 ± 0.6740	657763.2100 ± 235315.6482	100.0 %	0.0860 ± 0.0912	30118.4200 ± 32113.6083	1.2836 ± 0.3801	448790.0000 ± 133803.4600
Q-learning (π) ($\epsilon = 1$)	2.02001 ± 0.4996	685022.7200 ± 169670.6191	100.0 %	0.0539 ± 0.0569	18366.3400 ± 19499.6667	1.3425 ± 0.2770	455760.0000 ± 94888.6843
Model-free Dijkstra	0.00248 ± 0.0006	4423.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Model-free							
Asynchronous Value Iteration	0.04920 ± 0.0004	69570.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Model-free Value Iteration	0.05585 ± 0.0007	82594.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan

Algorithm (Problem 11)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)
Q-learning ($\epsilon = 0$)	0.11746 ± 0.0010	35316.0000 ± 0.0000	0.0 %	0.0052 ± 0.0001	1569.0000 ± 0.0000	0.0562 ± 0.0005	17000.0000 ± 0.0000
Q-learning ($\epsilon = 0.25$)	0.16348 ± 0.0017	48929.1500 ± 308.7902	0.0 %	0.0069 ± 0.0019	2083.6000 ± 575.2788	0.0674 ± 0.0018	20340.0000 ± 533.2917
Q-learning ($\epsilon = 0.5$)	0.24302 ± 0.0030	75008.7100 ± 538.2311	0.0 %	0.0083 ± 0.0028	2588.2400 ± 895.4427	0.0801 ± 0.0022	24950.0000 ± 683.7397
Q-learning ($\epsilon = 0.75$)	0.52179 ± 0.0074	166191.9100 ± 2022.8075	0.0 %	0.0099 ± 0.0041	3199.6400 ± 1346.4165	0.1070 ± 0.0044	34520.0000 ± 1403.4244
Q-learning ($\epsilon = 0.9$)	0.83570 ± 0.1904	272845.6200 ± 61283.9929	99.0 %	0.0093 ± 0.0061	3071.7600 ± 2019.3443	0.1416 ± 0.0071	46750.0000 ± 2329.6996
Q-learning ($\epsilon = 1$)	0.41468 ± 0.0594	144164.0500 ± 20481.0098	100.0 %	0.0054 ± 0.0041	1859.9000 ± 1424.8237	0.2162 ± 0.0341	75080.0000 ± 11847.0925
Q-learning (π) ($\epsilon = 1$)	0.42675 ± 0.0642	144673.6700 ± 21823.7557	100.0 %	0.0058 ± 0.0055	1936.9800 ± 1860.8880	0.2262 ± 0.0405	76650.0000 ± 13750.9091
Model-free Dijkstra	0.00248 ± 0.0005	4530.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Model-free							
Asynchronous Value Iteration	0.02155 ± 0.0005	30177.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Model-free Value Iteration	0.02846 ± 0.0005	41589.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan

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

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

Algorithm (Problem 14)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)
Q-learning ($\epsilon = 0$)	0.11687 ± 0.0011	36008.0000 ± 0.0000	0.0 %	0.0053 ± 0.0000	1615.0000 ± 0.0000	0.0517 ± 0.0005	16000.0000 ± 0.0000
Q-learning ($\epsilon = 0.25$)	0.15673 ± 0.0016	48378.9900 ± 210.8762	0.0 %	0.0078 ± 0.0023	2411.4400 ± 713.8361	0.0620 ± 0.0016	19250.0000 ± 433.0127
Q-learning ($\epsilon = 0.5$)	0.22881 ± 0.0030	72961.0200 ± 601.7355	0.0 %	0.0098 ± 0.0026	3142.1200 ± 846.0950	0.0729 ± 0.0020	23390.0000 ± 598.2474
Q-learning ($\epsilon = 0.75$)	0.47538 ± 0.0067	156573.9700 ± 1738.3132	0.0 %	0.0113 ± 0.0042	3729.0000 ± 1393.5140	0.0978 ± 0.0038	32440.0000 ± 1210.9500
Q-learning ($\epsilon = 0.9$)	0.48236 ± 0.0984	163299.2600 ± 32972.2192	100.0 %	0.0095 ± 0.0062	3224.5600 ± 2110.4774	0.1341 ± 0.0069	45590.0000 ± 2293.8832
Q-learning ($\epsilon = 1$)	0.30029 ± 0.0322	106857.6900 ± 11388.7065	100.0 %	0.0057 ± 0.0045	2018.8400 ± 1595.4904	0.1874 ± 0.0205	66620.0000 ± 7206.6358
Q-learning (π) ($\epsilon = 1$)	0.31998 ± 0.0378	108522.8800 ± 12668.4291	100.0 %	0.0054 ± 0.0050	1817.5800 ± 1713.0152	0.1942 ± 0.0222	65910.0000 ± 7482.1053
Model-free Dijkstra	0.00220 ± 0.0005	3711.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Model-free							
Asynchronous Value Iteration	0.02243 ± 0.0005	30808.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Model-free Value Iteration	0.02449 ± 0.0001	35112.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan

Algorithm (Problem 15)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)
Q-learning ($\epsilon = 0$)	0.11131 ± 0.0010	33372.0000 ± 0.0000	0.0 %	0.0040 ± 0.0000	1207.0000 ± 0.0000	0.0599 ± 0.0005	18000.0000 ± 0.0000
Q-learning ($\epsilon = 0.25$)	0.14548 ± 0.0015	43947.8600 ± 277.4724	0.0 %	0.0071 ± 0.0016	2168.5400 ± 506.3488	0.0703 ± 0.0018	21410.0000 ± 511.7617
Q-learning ($\epsilon = 0.5$)	0.20891 ± 0.0020	64789.3400 ± 537.2428	0.0 %	0.0081 ± 0.0020	2559.7000 ± 650.5667	0.0855 ± 0.0029	26800.0000 ± 894.4272
Q-learning ($\epsilon = 0.75$)	0.40536 ± 0.0057	130578.3200 ± 1350.2966	0.0 %	0.0085 ± 0.0036	2766.3800 ± 1169.2164	0.1189 ± 0.0062	38750.0000 ± 1986.8316
Q-learning ($\epsilon = 0.9$)	0.95809 ± 0.0194	315815.5100 ± 5887.7182	0.0 %	0.0065 ± 0.0040	2155.7000 ± 1334.3122	0.1624 ± 0.0131	54110.0000 ± 4402.0336
Q-learning ($\epsilon = 1$)	0.40327 ± 0.0676	141574.9600 ± 23520.4918	100.0 %	0.0037 ± 0.0030	1279.2800 ± 1055.8500	0.2170 ± 0.0238	76250.0000 ± 8462.1215
Q-learning (π) ($\epsilon = 1$)	0.39702 ± 0.0784	135803.2500 ± 26665.6822	100.0 %	0.0029 ± 0.0027	994.5000 ± 942.6673	0.2230 ± 0.0282	76310.0000 ± 9585.0874
Model-free Dijkstra	0.00253 ± 0.0007	4494.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Model-free							
Asynchronous Value Iteration	0.01561 ± 0.0002	22175.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Model-free Value Iteration	0.02710 ± 0.0005	39619.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan

Algorithm (Problem 16)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)
Q-learning ($\epsilon = 0$)	0.00285 ± 0.0000	1000.0000 ± 0.0000	100.0 %	0.0000 ± 0.0000	14.0000 ± 0.0000	0.0028 ± 0.0000	1000.0000 ± 0.0000
Q-learning ($\epsilon = 0.25$)	0.00289 ± 0.0000	1002.1600 ± 2.1294	100.0 %	0.0001 ± 0.0000	16.0000 ± 3.3941	0.0029 ± 0.0000	1000.0000 ± 0.0000
Q-learning ($\epsilon = 0.5$)	0.00283 ± 0.0000	1004.0000 ± 3.7094	100.0 %	0.0001 ± 0.0000	18.3000 ± 7.0050	0.0028 ± 0.0000	1000.0000 ± 0.0000
Q-learning ($\epsilon = 0.75$)	0.00278 ± 0.0000	1006.4000 ± 6.9685	100.0 %	0.0001 ± 0.0000	20.9800 ± 10.2781	0.0028 ± 0.0000	1000.0000 ± 0.0000
Q-learning ($\epsilon = 0.9$)	0.00276 ± 0.0000	1013.2800 ± 14.3095	100.0 %	0.0001 ± 0.0001	26.3400 ± 20.9496	0.0027 ± 0.0000	1000.0000 ± 0.0000
Q-learning ($\epsilon = 1$)	0.00266 ± 0.0001	1019.6000 ± 22.4410	100.0 %	0.0001 ± 0.0001	23.6000 ± 19.2603	0.0026 ± 0.0000	1000.0000 ± 0.0000
Q-learning (π) ($\epsilon = 1$)	0.00307 ± 0.0001	1023.6400 ± 24.1386	100.0 %	0.0001 ± 0.0001	21.6400 ± 22.2484	0.0030 ± 0.0000	1000.0000 ± 0.0000
Model-free Dijkstra	0.00003 ± 0.0000	28.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Model-free							
Asynchronous Value Iteration	0.00007 ± 0.0000	69.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Model-free Value Iteration	0.00008 ± 0.0000	79.0000 ± 0.0000	100.0 %	nan ± nan	nan ± nan	nan ± nan	nan ± nan

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

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

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

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

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

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

D Figures for stochastic experiments

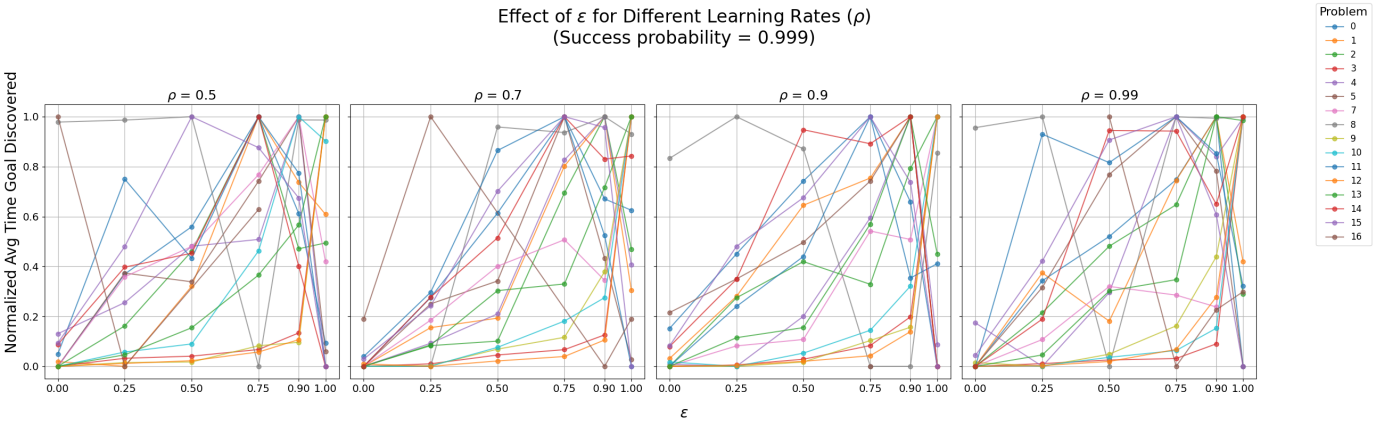

Fig. 6. The effect of decreasing greediness in a policy when applied on different problems with $\gamma = 0.999$ and varying learning rate ρ.

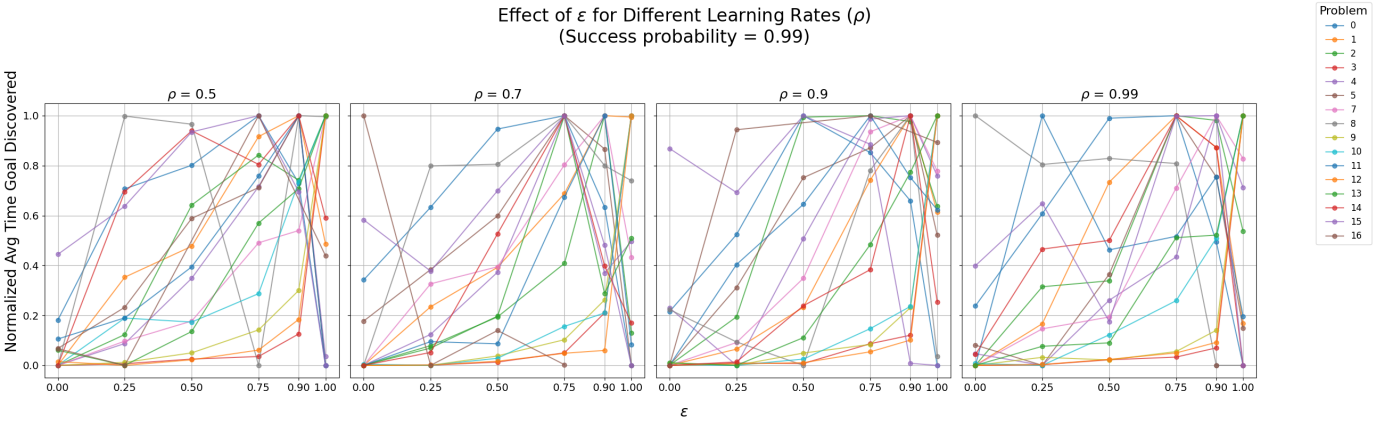

Fig. 7. The effect of decreasing greediness in a policy when applied on different problems with $\gamma = 0.99$ and varying learning rate ρ.

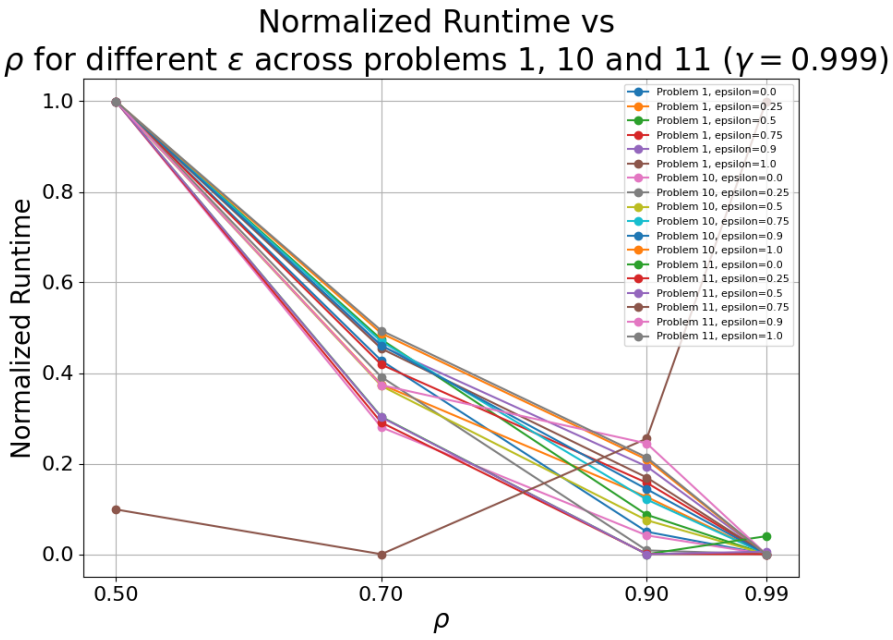

Fig. 8. As ρ increases, the run time for Q-learning decreases when $\gamma = 0.999$ for all ϵ with the exception of Problem 11 with $\epsilon = 0.75$.

E 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 ϵ-decay in Tables 22, 23 and 24. Additional experiments with low γ 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 ρ in stochastic problem with $\gamma = 0.99$ (Problem 1).

Algorithm (Problem 1) ($\gamma = 0.99$)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.44775 ± 0.0094	110018.9500 ± 120.6576	0.0%	0.0041 ± 0.0013	1016.9000 ± 309.0232	0.1815 ± 0.0082	44680.0000 ± 1907.2493	100.0%	28/34
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.42652 ± 0.0104	103655.8100 ± 848.3384	0.0%	0.0040 ± 0.0012	968.4300 ± 267.7274	0.1359 ± 0.0100	33040.0000 ± 2262.3881	100.0%	28/1456
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.41722 ± 0.0076	102035.4600 ± 1339.1680	0.0%	0.0042 ± 0.0013	1014.5900 ± 292.7621	0.1306 ± 0.0429	31970.0000 ± 10627.7514	100.0%	28/920
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.42034 ± 0.0090	102414.8500 ± 1556.8528	0.0%	0.0042 ± 0.0012	1026.1100 ± 282.5215	0.2078 ± 0.0750	50697.6744 ± 18153.7809	86.0%	28/2205
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.56923 ± 0.0061	141115.7100 ± 320.6419	0.0%	0.0059 ± 0.0015	1472.4300 ± 377.8420	0.3409 ± 0.0901	84917.6471 ± 22314.9708	85.0%	28/32
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.54557 ± 0.0069	134211.4300 ± 390.0833	0.0%	0.0050 ± 0.0014	1222.1600 ± 355.0452	0.3071 ± 0.1184	75800.0000 ± 28881.8282	50.0%	28/31
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	0.52808 ± 0.0066	130873.3600 ± 466.7216	0.0%	0.0052 ± 0.0015	1279.0600 ± 358.8929	0.2649 ± 0.1150	65640.0000 ± 28480.0000	25.0%	28/31
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	0.52894 ± 0.0130	130261.6100 ± 474.9027	0.0%	0.0054 ± 0.0016	1322.2200 ± 398.5655	0.3375 ± 0.0000	83000.0000 ± 0.0000	1.0%	28/31
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.82059 ± 0.0169	206637.7200 ± 779.9925	0.0%	0.0073 ± 0.0018	1864.0100 ± 450.2174	0.5143 ± 0.1522	129593.7500 ± 37137.2954	32.0%	28/31
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.79430 ± 0.0118	200809.2200 ± 926.3823	0.0%	0.0067 ± 0.0017	1694.9400 ± 416.3421	0.4310 ± 0.1809	109380.9524 ± 45658.6467	21.0%	28/31
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	0.78556 ± 0.0137	198195.8800 ± 936.3099	0.0%	0.0065 ± 0.0020	1626.9300 ± 505.6041	0.3918 ± 0.1863	100000.0000 ± 47588.5140	3.0%	28/31
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.77585 ± 0.0109	197337.6900 ± 857.3310	0.0%	0.0063 ± 0.0017	1622.1800 ± 433.5110	nan ± nan	nan ± nan	0.0%	28/1474
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	1.57260 ± 0.0235	413107.3800 ± 2630.4240	0.0%	0.0093 ± 0.0033	2458.8300 ± 861.9200	0.9108 ± 0.3088	240235.2941 ± 80468.8682	34.0%	28/31
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	1.56681 ± 0.0272	410248.2100 ± 3179.0668	0.0%	0.0087 ± 0.0035	2291.0500 ± 915.9052	0.5933 ± 0.3101	155533.3333 ± 81770.3016	15.0%	28/31
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	1.56187 ± 0.0176	411440.7600 ± 3119.8781	0.0%	0.0086 ± 0.0027	2261.1900 ± 731.0403	1.4754 ± 0.0000	391000.0000 ± 0.0000	1.0%	28/30
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	1.62392 ± 0.0210	422192.3300 ± 3032.9179	0.0%	0.0084 ± 0.0029	2187.4200 ± 769.7732	nan ± nan	nan ± nan	0.0%	28/31
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	3.82287 ± 0.1084	1019592.4000 ± 10171.1316	0.0%	0.0099 ± 0.0048	2652.3500 ± 1285.4855	2.0493 ± 0.9742	549081.6327 ± 260621.2231	49.0%	28/31
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	3.84191 ± 0.0587	1027729.5600 ± 11088.5495	0.0%	0.0085 ± 0.0048	2290.7600 ± 1295.1463	2.0426 ± 1.0165	547480.0000 ± 272473.1356	25.0%	28/31
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	4.00248 ± 0.0630	1081800.3000 ± 12373.4476	0.0%	0.0091 ± 0.0056	2471.7700 ± 1518.8372	2.8513 ± 0.9324	768333.3333 ± 246698.6466	3.0%	28/33
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	4.38596 ± 0.0606	1190762.1000 ± 13875.0949	0.0%	0.0092 ± 0.0045	2494.2500 ± 1240.8426	nan ± nan	nan ± nan	0.0%	28/532
Q-learning ($\rho = 0.5, \epsilon = 1$)	15.46274 ± 0.1914	4324185.7000 ± 48051.5290	0.0%	0.0053 ± 0.0040	1481.3700 ± 1126.3129	5.2842 ± 3.5571	1478922.2222 ± 997581.0992	90.0%	28/33
Q-learning ($\rho = 0.7, \epsilon = 1$)	15.38696 ± 0.2003	4319096.5300 ± 52853.4397	0.0%	0.0057 ± 0.0051	1588.3300 ± 1441.8374	5.6281 ± 4.1104	1579575.0000 ± 1155420.2458	80.0%	28/31
Q-learning ($\rho = 0.9, \epsilon = 1$)	15.39582 ± 0.2069	4319567.6700 ± 54011.2494	0.0%	0.0059 ± 0.0048	1638.9100 ± 1327.6757	7.1078 ± 4.6024	1995958.3333 ± 1293126.1372	24.0%	28/3346
Q-learning ($\rho = 0.99, \epsilon = 1$)	15.39603 ± 0.1846	4314091.5500 ± 48947.1783	0.0%	0.0060 ± 0.0048	1661.0400 ± 1336.9004	nan ± nan	nan ± nan	0.0%	28/32
Async Value Iteration	0.15656 ± 0.0014	137472.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	28/33
Value Iteration	0.22934 ± 0.0020	194752.0000 ± 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 ρ in stochastic problem with $\gamma = 0.999$ (Problem 1).

Algorithm (Problem 1) ($\gamma = 0.999$)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.45780 ± 0.0072	108128.1300 ± 66.4493	0.0%	0.0046 ± 0.0004	1072.3100 ± 94.0911	0.1763 ± 0.0038	41650.0000 ± 476.9696	100.0%	28/7264
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.42858 ± 0.0087	100704.8600 ± 108.2371	0.0%	0.0045 ± 0.0005	1053.1400 ± 123.3148	0.1273 ± 0.0027	29950.0000 ± 217.9449	100.0%	28/1456
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.40930 ± 0.0046	96842.1900 ± 116.4245	0.0%	0.0044 ± 0.0006	1041.3000 ± 148.1313	0.0973 ± 0.0012	23020.0000 ± 140.0000	100.0%	28/30
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.40677 ± 0.0041	96121.0500 ± 195.6695	0.0%	0.0044 ± 0.0006	1044.4500 ± 130.2082	0.0888 ± 0.0008	21000.0000 ± 0.0000	100.0%	28/32
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.57845 ± 0.0051	138790.8900 ± 386.8301	0.0%	0.0061 ± 0.0017	1468.7500 ± 399.6031	0.1898 ± 0.0019	45920.0000 ± 305.9412	100.0%	28/30
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.54797 ± 0.0058	131788.1800 ± 402.8578	0.0%	0.0050 ± 0.0014	1217.3300 ± 349.6271	0.1381 ± 0.0026	33490.0000 ± 519.5190	100.0%	28/30
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	0.53586 ± 0.0064	128248.1100 ± 438.3435	0.0%	0.0051 ± 0.0015	1239.5400 ± 340.2047	0.1100 ± 0.0024	26560.0000 ± 496.3869	100.0%	28/29
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	0.52970 ± 0.0064	127530.0700 ± 370.5002	0.0%	0.0053 ± 0.0016	1278.4100 ± 378.5350	0.1006 ± 0.0024	24320.0000 ± 466.4762	100.0%	28/29
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.81914 ± 0.0101	203280.6600 ± 1146.9714	0.0%	0.0073 ± 0.0020	1814.4300 ± 480.8558	0.2137 ± 0.0035	53480.0000 ± 556.4171	100.0%	28/30
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.78980 ± 0.0079	197246.0900 ± 978.0357	0.0%	0.0068 ± 0.0018	1702.0100 ± 435.4708	0.1581 ± 0.0025	39800.0000 ± 489.8979	100.0%	28/31
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	0.76821 ± 0.0160	196367.9900 ± 1204.9415	0.0%	0.0064 ± 0.0020	1632.6200 ± 507.3923	0.1254 ± 0.0032	32350.0000 ± 606.2178	100.0%	28/30
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.76335 ± 0.0084	197826.2100 ± 875.8254	0.0%	0.0063 ± 0.0016	1634.3100 ± 421.2690	0.1144 ± 0.0024	29800.0000 ± 529.1503	100.0%	28/30
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	1.20039 ± 0.2329	321255.2700 ± 62103.2471	88.0%	0.0088 ± 0.0033	2362.1600 ± 872.5669	0.2545 ± 0.0053	68620.0000 ± 1172.8598	100.0%	28/29
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	0.80764 ± 0.1909	215408.4700 ± 50762.6464	100.0%	0.0083 ± 0.0031	2217.5700 ± 840.3794	0.1969 ± 0.0063	52760.0000 ± 1304.7605	100.0%	28/29
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	0.63141 ± 0.1970	169741.5500 ± 52725.1623	100.0%	0.0091 ± 0.0037	2441.8600 ± 986.8430	0.1640 ± 0.0048	44210.0000 ± 1194.1105	100.0%	28/29
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	0.52489 ± 0.1930	140000.4800 ± 50657.7332	100.0%	0.0090 ± 0.0034	2408.0900 ± 894.9593	0.1516 ± 0.0045	40480.0000 ± 1081.4805	100.0%	28/30
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	0.51857 ± 0.0569	140998.0100 ± 15374.5296	100.0%	0.0093 ± 0.0056	2524.0900 ± 1527.4172	0.3294 ± 0.0125	89500.0000 ± 3185.9065	100.0%	28/30
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	0.38885 ± 0.0494	106196.5900 ± 13726.0244	100.0%	0.0095 ± 0.0053	2586.7900 ± 1449.9621	0.2606 ± 0.0111	71060.0000 ± 3012.7064	100.0%	28/30
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	0.32507 ± 0.0361	89175.6700 ± 9912.0126	100.0%	0.0084 ± 0.0049	2299.5000 ± 1333.3552	0.2203 ± 0.0109	60350.0000 ± 2878.8018	100.0%	28/29
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	0.27856 ± 0.0271	75963.0100 ± 7419.7872	100.0%	0.0089 ± 0.0046	2421.2300 ± 1268.8343	0.2044 ± 0.0109	55670.0000 ± 2915.6646	100.0%	28/29
Q-learning ($\rho = 0.5, \epsilon = 1$)	0.68562 ± 0.0786	190240.5100 ± 21265.5439	100.0%	0.0063 ± 0.0062	1735.3900 ± 1720.3775	0.6086 ± 0.0661	168960.0000 ± 18121.2141	100.0%	28/29
Q-learning ($\rho = 0.7, \epsilon = 1$)	0.53277 ± 0.0642	148527.9800 ± 17641.4631	100.0%	0.0053 ± 0.0047	1466.7500 ± 1296.5192	0.4704 ± 0.0622	131110.0000 ± 17159.1929	100.0%	28/30
Q-learning ($\rho = 0.9, \epsilon = 1$)	0.45312 ± 0.0586	126285.1000 ± 16381.5359	100.0%	0.0064 ± 0.0048	1770.8200 ± 1325.8271	0.3915 ± 0.0569	109050.0000 ± 15879.7827	100.0%	28/30
Q-learning ($\rho = 0.99, \epsilon = 1$)	0.40572 ± 0.0491	113338.4300 ± 13755.4512	100.0%	0.0064 ± 0.0059	1768.5900 ± 1659.4249	0.3512 ± 0.0450	98050.0000 ± 12561.3494	100.0%	28/30
Async Value Iteration	0.12869 ± 0.0009	114560.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	28/29
Value Iteration	0.17870 ± 0.0024	154656.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	28/30

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

Algorithm (Problem 10) ($\gamma = 0.99$)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	1.09601 ± 0.0226	270673.0900 ± 1030.1652	0.0%	0.0098 ± 0.0041	2379.9400 ± 996.5445	0.4524 ± 0.0133	110860.0000 ± 2561.3278	100.0%	64/322
Q-learning ($\rho = 0.7, \epsilon = 0$)	1.03734 ± 0.0188	256745.5500 ± 3275.4224	0.0%	0.0099 ± 0.0039	2416.5100 ± 938.7294	0.3480 ± 0.1000	85742.2680 ± 25179.4706	97.0%	64/1614
Q-learning ($\rho = 0.9, \epsilon = 0$)	1.05040 ± 0.0185	259065.8600 ± 3638.1183	0.0%	0.0092 ± 0.0033	2244.6800 ± 804.8872	0.3135 ± 0.1367	77466.6667 ± 33639.9894	60.0%	66/45219
Q-learning ($\rho = 0.99, \epsilon = 0$)	1.07131 ± 0.0178	262565.8200 ± 3765.1413	0.0%	0.0093 ± 0.0036	2259.5400 ± 896.9925	0.2651 ± 0.0000	65000.0000 ± 0.0000	1.0%	66/2596
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	1.43981 ± 0.0548	354087.5500 ± 622.1560	0.0%	0.0140 ± 0.0077	3431.0800 ± 1861.5218	0.6260 ± 0.1351	154170.0000 ± 32728.2920	100.0%	64/70
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	1.34549 ± 0.0105	334650.5700 ± 1071.9651	0.0%	0.0105 ± 0.0067	2620.3600 ± 1668.5465	0.4836 ± 0.1409	120480.0000 ± 35038.4018	100.0%	64/112
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	1.34040 ± 0.1247	328920.8100 ± 1294.5546	0.0%	0.0118 ± 0.0073	2921.1600 ± 1811.3192	0.3797 ± 0.1185	93410.0000 ± 27494.7613	100.0%	64/7336
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	1.31547 ± 0.0130	328258.1500 ± 1489.9179	0.0%	0.0098 ± 0.0060	2448.4600 ± 1519.3635	0.5097 ± 0.2413	127547.3684 ± 60110.9095	95.0%	64/1706
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	2.08346 ± 0.0157	531713.7700 ± 1337.2434	0.0%	0.0181 ± 0.0094	4630.6000 ± 2360.6799	0.7959 ± 0.2102	203550.0000 ± 53193.3031	100.0%	64/68
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	2.01877 ± 0.0167	514318.0800 ± 1448.9041	0.0%	0.0148 ± 0.0080	3775.6000 ± 2037.4858	0.6919 ± 0.2492	176790.0000 ± 63480.7522	100.0%	64/79
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	2.00751 ± 0.0192	508357.3700 ± 1698.5668	0.0%	0.0158 ± 0.0096	4033.5500 ± 2437.1517	0.5771 ± 0.2580	146590.0000 ± 64556.9663	100.0%	64/795
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	1.99043 ± 0.0217	508076.7500 ± 1609.1026	0.0%	0.0153 ± 0.0081	3904.9100 ± 2061.3847	0.6288 ± 0.3471	160909.0909 ± 88437.5133	99.0%	64/3689
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	4.16759 ± 0.0386	1099049.3400 ± 5207.7051	0.0%	0.0253 ± 0.0135	6721.1800 ± 3592.8100	1.0769 ± 0.3966	285020.0000 ± 104270.9912	100.0%	64/68
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	4.12914 ± 0.0413	1091002.3000 ± 4279.7067	0.0%	0.0226 ± 0.0131	6016.9800 ± 3466.9508	0.9614 ± 0.4025	254810.0000 ± 105862.1457	100.0%	64/68
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	4.12844 ± 0.0431	1091267.5600 ± 4694.5104	0.0%	0.0232 ± 0.0132	6188.1300 ± 3511.1027	0.8190 ± 0.4050	217540.0000 ± 106509.2879	100.0%	64/654
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	4.16205 ± 0.0467	1093979.6700 ± 4652.0697	0.0%	0.0225 ± 0.0137	5966.1100 ± 3618.1097	1.0492 ± 0.7983	276909.0909 ± 210896.2640	99.0%	64/1395
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	10.57651 ± 0.1140	2868949.5500 ± 18988.1324	0.0%	0.0421 ± 0.0250	11472.2500 ± 6795.5047	1.8967 ± 1.0263	514400.0000 ± 278406.6450	100.0%	64/69
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	10.54571 ± 0.1132	2879704.9100 ± 18072.6930	0.0%	0.0376 ± 0.0232	10316.3800 ± 6390.1483	1.7248 ± 0.8627	471640.0000 ± 234910.3455	100.0%	64/323
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	10.65997 ± 0.1173	2907657.8700 ± 20629.3163	0.0%	0.0408 ± 0.0238	11182.0800 ± 6486.2508	1.5113 ± 0.9102	413190.0000 ± 248451.4719	100.0%	64/1372
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	10.66764 ± 0.1191	2913088.7500 ± 18333.2363	0.0%	0.0362 ± 0.0250	9947.0600 ± 6859.7258	1.8766 ± 1.3522	512730.0000 ± 367422.6954	100.0%	64/1052
Q-learning ($\rho = 0.5, \epsilon = 1$)	30.93903 ± 0.2059	8716611.8600 ± 19277.6545	0.0%	0.1126 ± 0.1084	31658.2200 ± 30581.5958	3.5574 ± 1.1094	999950.0000 ± 313103.3176	100.0%	64/70
Q-learning ($\rho = 0.7, \epsilon = 1$)	30.89782 ± 0.2076	8711546.9300 ± 17111.6194	0.0%	0.0979 ± 0.0971	27449.0200 ± 27265.8711	2.6415 ± 0.7523	741970.0000 ± 211201.6314	100.0%	64/68
Q-learning ($\rho = 0.9, \epsilon = 1$)	30.84969 ± 0.1420	8716302.9200 ± 16471.9515	0.0%	0.0793 ± 0.0704	22381.9100 ± 20016.9771	2.4499 ± 0.9134	690910.0000 ± 257280.1623	100.0%	64/510
Q-learning ($\rho = 0.99, \epsilon = 1$)	30.94423 ± 0.1896	8716436.7700 ± 18109.7792	0.0%	0.1120 ± 0.1162	31433.8200 ± 32655.3818	2.4456 ± 1.0780	687030.0000 ± 303158.6533	100.0%	64/610
Async Value Iteration	0.24640 ± 0.0036	212106.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	64/69
Value Iteration	0.40702 ± 0.0028	344288.0000 ± 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 ρ in stochastic problem with $\gamma = 0.999$ (Problem 10).

Algorithm (Problem 10) ($\gamma = 0.999$)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	1.06396 ± 0.0155	265867.3800 ± 75.5957	0.0%	0.0118 ± 0.0029	2910.3800 ± 728.7117	0.4371 ± 0.0062	108280.0000 ± 601.3319	100.0%	64/48894
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.97940 ± 0.0126	245376.2300 ± 1468.6347	0.0%	0.0101 ± 0.0016	2489.1400 ± 385.4470	0.3118 ± 0.0046	77440.0000 ± 682.9348	100.0%	64/8716
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.95135 ± 0.0260	237506.9900 ± 2795.6508	0.0%	0.0102 ± 0.0015	2513.4300 ± 374.4821	0.2440 ± 0.0167	60840.0000 ± 4251.3998	100.0%	64/100002
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.94648 ± 0.0174	236133.6700 ± 2755.9321	0.0%	0.0099 ± 0.0017	2432.8900 ± 408.7694	0.2290 ± 0.0314	57000.0000 ± 7900.6329	100.0%	64/27731
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	1.39075 ± 0.0168	348458.2300 ± 550.3162	0.0%	0.0113 ± 0.0067	2844.8200 ± 1667.9490	0.4796 ± 0.0074	120530.0000 ± 921.4662	100.0%	64/66
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	1.30497 ± 0.0156	328004.3100 ± 553.7289	0.0%	0.0109 ± 0.0067	2743.1800 ± 1696.3778	0.3480 ± 0.0050	87900.0000 ± 754.9834	100.0%	64/66
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	1.25100 ± 0.0116	317322.7300 ± 578.4866	0.0%	0.0097 ± 0.0054	2449.3400 ± 1376.1401	0.2732 ± 0.0035	69630.0000 ± 673.1270	100.0%	64/66
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	1.24982 ± 0.0203	314602.3800 ± 649.2282	0.0%	0.0104 ± 0.0063	2622.7000 ± 1574.8014	0.2513 ± 0.0056	63510.0000 ± 754.9172	100.0%	64/66
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	2.01510 ± 0.0174	523541.2100 ± 1416.2376	0.0%	0.0165 ± 0.0087	4313.0500 ± 2250.0812	0.5370 ± 0.0081	140430.0000 ± 1544.3769	100.0%	64/66
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	1.95291 ± 0.0191	506799.5600 ± 1395.7975	0.0%	0.0145 ± 0.0084	3800.7300 ± 2189.7840	0.4021 ± 0.0068	105050.0000 ± 1283.5498	100.0%	64/66
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	1.92356 ± 0.0206	498905.5000 ± 1280.5515	0.0%	0.0134 ± 0.0077	3502.3900 ± 2011.4538	0.3275 ± 0.0065	85460.0000 ± 1169.7863	100.0%	64/66
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	1.91614 ± 0.0188	497578.6300 ± 1433.7520	0.0%	0.0147 ± 0.0070	3830.3700 ± 1822.5697	0.3029 ± 0.0058	79100.0000 ± 1268.8578	100.0%	64/67
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	1.79071 ± 0.3039	479534.0000 ± 80679.4280	100.0%	0.0234 ± 0.0129	6311.8400 ± 3458.1080	0.6708 ± 0.0162	180730.0000 ± 3709.0565	100.0%	64/66
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	1.35937 ± 0.2328	364887.2700 ± 62023.3435	100.0%	0.0243 ± 0.0137	6523.8400 ± 3672.1303	0.5161 ± 0.0099	139210.0000 ± 2173.9135	100.0%	64/66
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	1.07579 ± 0.2083	286890.2300 ± 54993.0725	100.0%	0.0243 ± 0.0130	6500.8100 ± 3462.0776	0.4383 ± 0.0111	117390.0000 ± 2603.4400	100.0%	64/66
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	0.97765 ± 0.2121	262076.9300 ± 56073.2616	100.0%	0.0251 ± 0.0131	6727.4000 ± 3498.1394	0.4122 ± 0.0107	110750.0000 ± 2511.4737	100.0%	64/66
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	1.04660 ± 0.0937	284588.2100 ± 24490.8895	100.0%	0.0357 ± 0.0210	9731.1100 ± 5716.6608	0.8668 ± 0.0338	235920.0000 ± 7632.4046	100.0%	64/66
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	0.83691 ± 0.0952	229701.4300 ± 25612.8380	100.0%	0.0368 ± 0.0228	10096.0300 ± 6266.5153	0.6819 ± 0.0236	187190.0000 ± 6042.6732	100.0%	64/66
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	0.71331 ± 0.0724	194477.8400 ± 19500.5225	100.0%	0.0375 ± 0.0232	10217.5800 ± 6302.4782	0.5952 ± 0.0212	162210.0000 ± 5532.2599	100.0%	64/67
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	0.65742 ± 0.0572	180012.3600 ± 15260.4553	100.0%	0.0371 ± 0.0219	10147.8100 ± 6017.0427	0.5611 ± 0.0225	153560.0000 ± 5817.7659	100.0%	64/66
Q-learning ($\rho = 0.5, \epsilon = 1$)	4.84662 ± 1.4434	1342809.3100 ± 399317.5826	100.0%	0.1142 ± 0.1162	31313.3500 ± 31759.6827	2.9789 ± 0.5718	824560.0000 ± 157018.8091	100.0%	64/66
Q-learning ($\rho = 0.7, \epsilon = 1$)	3.52291 ± 0.9998	985856.3500 ± 282047.3120	100.0%	0.0975 ± 0.0970	27210.7400 ± 27062.4015	2.2450 ± 0.4430	627880.0000 ± 124366.2559	100.0%	64/68
Q-learning ($\rho = 0.9, \epsilon = 1$)	2.80282 ± 1.0162	777332.6400 ± 281713.9771	100.0%	0.0900 ± 0.0809	24856.4400 ± 22392.3595	1.8761 ± 0.4569	519830.0000 ± 125921.3290	100.0%	64/67
Q-learning ($\rho = 0.99, \epsilon = 1$)	2.26297 ± 0.7142	623074.5400 ± 199075.3965	100.0%	0.0972 ± 0.0793	26692.6900 ± 21712.2338	1.6362 ± 0.4014	449470.0000 ± 110564.1402	100.0%	64/66
Async Value Iteration	0.21926 ± 0.0162	178292.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	64/66
Value Iteration	0.35463 ± 0.0252	288956.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	64/66

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

Algorithm (Problem 11) ($\gamma = 0.99$)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.39799 ± 0.0210	95423.6800 ± 567.6079	0.0%	0.0067 ± 0.0011	1633.4300 ± 258.5025	0.1518 ± 0.0101	36850.0000 ± 2321.0989	100.0%	22/100002
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.37312 ± 0.0115	90568.8000 ± 2163.3390	0.0%	0.0067 ± 0.0011	1637.7600 ± 266.1527	0.1241 ± 0.0323	30404.0404 ± 7841.7675	99.0%	22/310
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.38073 ± 0.0124	92259.2500 ± 2742.0887	0.0%	0.0068 ± 0.0010	1675.1000 ± 256.6727	0.1323 ± 0.0711	32383.5616 ± 17337.6067	73.0%	22/3046
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.38594 ± 0.0130	93504.7500 ± 2852.7542	0.0%	0.0065 ± 0.0008	1590.2000 ± 183.6596	0.1885 ± 0.0765	45947.3684 ± 18453.2801	38.0%	22/696
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.51791 ± 0.0067	126854.4000 ± 347.8654	0.0%	0.0080 ± 0.0018	1994.0800 ± 454.6916	0.1593 ± 0.0079	39700.0000 ± 1920.9373	100.0%	22/27
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.49021 ± 0.0080	119017.3200 ± 575.3976	0.0%	0.0074 ± 0.0015	1833.9600 ± 376.3915	0.1308 ± 0.0220	32300.0000 ± 5243.0907	100.0%	22/26
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	0.47611 ± 0.0072	115911.2900 ± 736.9714	0.0%	0.0075 ± 0.0017	1864.9900 ± 422.5565	0.1425 ± 0.0551	35121.2121 ± 13224.0008	99.0%	22/82
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	0.47157 ± 0.0060	115410.4400 ± 700.0474	0.0%	0.0074 ± 0.0017	1838.1700 ± 438.6807	0.2307 ± 0.1067	56633.3333 ± 25843.7398	30.0%	22/2063
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.76524 ± 0.0090	192499.1600 ± 812.2625	0.0%	0.0083 ± 0.0030	2131.3800 ± 781.6668	0.1784 ± 0.0105	45770.0000 ± 2591.7369	100.0%	22/25
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.73667 ± 0.0096	185125.1100 ± 827.9211	0.0%	0.0083 ± 0.0028	2118.1700 ± 725.6840	0.1429 ± 0.0185	36590.0000 ± 4630.5399	100.0%	22/25
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	0.72397 ± 0.0089	181844.3900 ± 970.7109	0.0%	0.0083 ± 0.0028	2120.4300 ± 740.6331	0.1584 ± 0.0545	40360.0000 ± 13635.6298	100.0%	22/1680
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.72236 ± 0.0078	181875.7900 ± 944.6629	0.0%	0.0088 ± 0.0025	2245.8600 ± 653.2801	0.3493 ± 0.1707	88333.3333 ± 42697.5335	21.0%	22/2313
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	1.55484 ± 0.0213	404114.4100 ± 2614.4949	0.0%	0.0112 ± 0.0039	2974.8200 ± 1040.4007	0.2188 ± 0.0166	57760.0000 ± 4393.4497	100.0%	22/25
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	1.53926 ± 0.0216	400345.6600 ± 3194.0362	0.0%	0.0099 ± 0.0047	2622.3500 ± 1254.2456	0.1835 ± 0.0228	48470.0000 ± 6104.8423	100.0%	22/25
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	1.54499 ± 0.0206	402185.0100 ± 3280.6152	0.0%	0.0107 ± 0.0040	2819.8400 ± 1073.0321	0.1793 ± 0.0451	47290.0000 ± 11888.8982	100.0%	22/25
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	1.57677 ± 0.0174	411567.1500 ± 2845.5532	0.0%	0.0101 ± 0.0040	2689.6600 ± 1072.7643	0.7294 ± 0.4547	190827.5862 ± 118468.1156	29.0%	22/25
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	4.06947 ± 0.0661	1089811.3000 ± 11695.3669	0.0%	0.0101 ± 0.0056	2707.5800 ± 1508.7473	0.2907 ± 0.0359	78120.0000 ± 9666.7264	100.0%	22/24
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	4.13515 ± 0.0593	1091747.8400 ± 10810.2072	0.0%	0.0104 ± 0.0059	2770.8300 ± 1585.6226	0.2434 ± 0.0409	65000.0000 ± 10853.5708	100.0%	22/25
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	4.26611 ± 0.0819	1124205.1400 ± 12992.2186	0.0%	0.0097 ± 0.0055	2599.7100 ± 1492.5060	0.2830 ± 0.0883	75330.0000 ± 23466.1693	100.0%	22/3090
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	4.51425 ± 0.0887	1198876.9800 ± 14514.8737	0.0%	0.0089 ± 0.0053	2387.6900 ± 1443.7231	2.0707 ± 1.2129	548638.2979 ± 320207.8066	47.0%	22/308
Q-learning ($\rho = 0.5, \epsilon = 1$)	17.38310 ± 0.2022	4830226.6800 ± 48616.3395	0.0%	0.0078 ± 0.0067	2157.7200 ± 1850.2128	0.4759 ± 0.0839	132370.0000 ± 23297.4913	100.0%	22/26
Q-learning ($\rho = 0.7, \epsilon = 1$)	17.41050 ± 0.2441	4828188.3500 ± 55089.5831	0.0%	0.0073 ± 0.0064	2009.8400 ± 1783.1239	0.4094 ± 0.1053	113350.0000 ± 29066.6045	100.0%	22/25
Q-learning ($\rho = 0.9, \epsilon = 1$)	17.54003 ± 0.2390	4841830.8000 ± 54715.3973	0.0%	0.0066 ± 0.0056	1799.5900 ± 1538.6349	0.4463 ± 0.1492	123170.0000 ± 41225.4909	100.0%	22/26
Q-learning ($\rho = 0.99, \epsilon = 1$)	17.47704 ± 0.2397	4828782.2700 ± 54208.8381	0.0%	0.0064 ± 0.0053	1769.1100 ± 1467.0934	7.3585 ± 4.5320	2033564.5161 ± 1253221.1480	62.0%	22/26
Async Value Iteration	0.20147 ± 0.0022	178048.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	22/25
Value Iteration	0.27400 ± 0.0051	232832.0000 ± 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 ρ in stochastic problem with $\gamma = 0.999$ (Problem 11).

Algorithm (Problem 11) ($\gamma = 0.999$)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.38377 ± 0.0045	93082.8800 ± 193.1769	0.0%	0.0056 ± 0.0004	1375.0100 ± 86.5073	0.1436 ± 0.0018	35020.0000 ± 140.0000	100.0%	22/26
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.35480 ± 0.0113	85400.1200 ± 660.9158	0.0%	0.0056 ± 0.0005	1372.6300 ± 108.3105	0.1035 ± 0.0017	25130.0000 ± 364.8287	100.0%	22/46025
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.34218 ± 0.0090	82557.8800 ± 1606.2870	0.0%	0.0056 ± 0.0003	1357.6200 ± 65.3582	0.0828 ± 0.0021	20160.0000 ± 417.6123	100.0%	22/21524
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.34383 ± 0.0101	82572.4900 ± 1944.7986	0.0%	0.0056 ± 0.0003	1360.0700 ± 77.6709	0.0793 ± 0.0028	19240.0000 ± 649.9231	100.0%	22/8643
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.51539 ± 0.0073	124786.3700 ± 373.1031	0.0%	0.0080 ± 0.0019	1971.3500 ± 464.3022	0.1532 ± 0.0029	37780.0000 ± 414.2463	100.0%	22/24
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.47911 ± 0.0046	116563.4500 ± 358.9886	0.0%	0.0075 ± 0.0016	1861.1700 ± 391.1360	0.1129 ± 0.0013	28010.0000 ± 223.3831	100.0%	22/24
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	0.46427 ± 0.0052	112306.8800 ± 354.4297	0.0%	0.0077 ± 0.0013	1895.2000 ± 325.2237	0.0925 ± 0.0021	22780.0000 ± 481.2484	100.0%	22/23
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	0.46423 ± 0.0088	111698.8200 ± 385.9230	0.0%	0.0075 ± 0.0015	1836.5800 ± 373.8440	0.0872 ± 0.0028	21370.0000 ± 594.2222	100.0%	22/24
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.76484 ± 0.0100	189923.3500 ± 1146.3437	0.0%	0.0086 ± 0.0024	2169.1000 ± 609.7160	0.1696 ± 0.0030	43010.0000 ± 538.4236	100.0%	22/22
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.73026 ± 0.0113	183074.4400 ± 1037.7751	0.0%	0.0081 ± 0.0027	2074.3600 ± 687.7962	0.1280 ± 0.0031	32700.0000 ± 591.6080	100.0%	22/24
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	0.71527 ± 0.0085	180180.1500 ± 1102.2192	0.0%	0.0080 ± 0.0023	2046.0700 ± 605.3878	0.1064 ± 0.0027	27270.0000 ± 597.5784	100.0%	22/22
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.71556 ± 0.0084	181054.6100 ± 955.8296	0.0%	0.0083 ± 0.0022	2130.3800 ± 580.1694	0.0993 ± 0.0029	25570.0000 ± 696.4912	100.0%	22/23
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	1.58401 ± 0.0221	415905.4000 ± 5266.6597	0.0%	0.0103 ± 0.0038	2729.0300 ± 1019.6230	0.2032 ± 0.0054	54160.0000 ± 1293.9861	100.0%	22/23
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	1.57483 ± 0.0245	409969.3600 ± 4411.5591	0.0%	0.0095 ± 0.0041	2505.6300 ± 1089.8314	0.1598 ± 0.0045	42230.0000 ± 1156.3304	100.0%	22/24
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	1.59856 ± 0.0268	417749.4400 ± 5188.3083	0.0%	0.0109 ± 0.0040	2883.3600 ± 1064.7107	0.1383 ± 0.0054	36210.0000 ± 1351.2587	100.0%	22/24
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	1.66779 ± 0.0268	435555.6000 ± 3416.3876	0.0%	0.0106 ± 0.0038	2815.4700 ± 1007.1658	0.1289 ± 0.0040	34110.0000 ± 998.9494	100.0%	22/23
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	1.00714 ± 0.1571	269682.8100 ± 41926.1143	100.0%	0.0100 ± 0.0060	2708.0700 ± 1615.6574	0.2630 ± 0.0117	70880.0000 ± 2997.5990	100.0%	22/24
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	0.75531 ± 0.1075	200773.1600 ± 28429.8556	100.0%	0.0102 ± 0.0058	2714.7700 ± 1565.0885	0.2099 ± 0.0123	56100.0000 ± 3106.4449	100.0%	22/24
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	0.70365 ± 0.1405	184876.9200 ± 36768.9518	100.0%	0.0098 ± 0.0056	2589.9500 ± 1502.0573	0.1851 ± 0.0108	48920.0000 ± 2777.3369	100.0%	22/24
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	0.60536 ± 0.1262	160644.1800 ± 32985.1712	100.0%	0.0108 ± 0.0061	2890.6800 ± 1640.6114	0.1714 ± 0.0078	45720.0000 ± 2040.0000	100.0%	22/24
Q-learning ($\rho = 0.5, \epsilon = 1$)	0.87511 ± 0.1260	241990.0900 ± 34516.0792	100.0%	0.0075 ± 0.0064	2077.2500 ± 1780.5063	0.4522 ± 0.0631	125150.0000 ± 17723.0782	100.0%	22/23
Q-learning ($\rho = 0.7, \epsilon = 1$)	0.70148 ± 0.0778	193154.4900 ± 21280.4327	100.0%	0.0073 ± 0.0061	2000.0100 ± 1683.8940	0.3587 ± 0.0626	98660.0000 ± 17117.9555	100.0%	22/24
Q-learning ($\rho = 0.9, \epsilon = 1$)	0.60545 ± 0.0850	167366.2800 ± 23265.3554	100.0%	0.0072 ± 0.0059	1964.9300 ± 1621.4139	0.3087 ± 0.0572	85250.0000 ± 15863.4013	100.0%	22/22
Q-learning ($\rho = 0.99, \epsilon = 1$)	0.53213 ± 0.0721	148307.8100 ± 20421.1873	100.0%	0.0064 ± 0.0051	1779.6300 ± 1419.8986	0.2641 ± 0.0442	73390.0000 ± 12284.0506	100.0%	22/23
Async Value Iteration	0.16458 ± 0.0020	147232.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	22/23
Value Iteration	0.21802 ± 0.0008	184896.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	22/23

Table 13. Performance of DP methods and Q-learning as we change the learning rate ρ in stochastic problem with $\gamma = 0.5$ (Problem 1).

Algorithm (Problem 1) ($\gamma = 0.5$)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.05, \epsilon = 0$)	2.82370 ± 0.0630	829360.5000 ± 3538.3784	0.0%	0.0077 ± 0.0039	2314.0000 ± 1165.9239	nan ± nan	nan ± nan	0.0%	66/226
Q-learning ($\rho = 0.1, \epsilon = 0$)	1.93985 ± 0.0274	571004.0000 ± 4104.9053	0.0%	0.0092 ± 0.0051	2723.5000 ± 1523.5988	1.9076 ± 0.0383	562833.3333 ± 5698.4403	60.0%	67/152
Q-learning ($\rho = 0.3, \epsilon = 0$)	1.22874 ± 0.0288	362432.8000 ± 4119.4320	0.0%	0.0067 ± 0.0020	1878.2000 ± 560.1846	0.6407 ± 0.0242	189000.0000 ± 7886.6977	100.0%	50/99
Q-learning ($\rho = 0.5, \epsilon = 0$)	1.10654 ± 0.0330	326328.2000 ± 7451.1840	0.0%	0.0078 ± 0.0023	2399.8000 ± 831.8626	0.5686 ± 0.1974	169200.0000 ± 58245.6865	100.0%	49/109
Q-learning ($\rho = 0.7, \epsilon = 0$)	1.10783 ± 0.0419	326296.3000 ± 9296.8325	0.0%	0.0115 ± 0.0060	3333.8000 ± 1718.8984	0.3924 ± 0.1783	116500.0000 ± 52913.6088	100.0%	52/130
Q-learning ($\rho = 0.05, \epsilon = 0.25$)	2.82281 ± 0.0287	850959.0000 ± 3373.9975	0.0%	0.0077 ± 0.0055	2375.4000 ± 1552.5524	nan ± nan	nan ± nan	0.0%	54/170
Q-learning ($\rho = 0.1, \epsilon = 0.25$)	2.02926 ± 0.0190	612104.8000 ± 3219.9006	0.0%	0.0080 ± 0.0039	2455.9000 ± 1053.3333	2.0377 ± 0.0142	614000.0000 ± 3000.0000	20.0%	57/99
Q-learning ($\rho = 0.3, \epsilon = 0.25$)	1.54693 ± 0.0130	470008.1000 ± 3552.6818	0.0%	0.0087 ± 0.0045	2732.8000 ± 1307.4142	0.7149 ± 0.0292	217900.0000 ± 9449.3386	100.0%	54/222
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	1.68989 ± 0.0212	509964.5000 ± 2610.0875	0.0%	0.0105 ± 0.0046	2987.6000 ± 1335.1519	0.5211 ± 0.0686	156400.0000 ± 21228.2830	100.0%	74/229
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	2.01671 ± 0.0204	609987.5000 ± 5718.5885	0.0%	0.0107 ± 0.0056	3137.0000 ± 1482.0965	0.3924 ± 0.0668	119900.0000 ± 21139.7729	100.0%	56/283
Q-learning ($\rho = 0.05, \epsilon = 0.5$)	3.01899 ± 0.0342	930272.2000 ± 3023.9536	0.0%	0.0079 ± 0.0033	2514.2000 ± 1067.4638	nan ± nan	nan ± nan	0.0%	57/127
Q-learning ($\rho = 0.1, \epsilon = 0.5$)	2.35263 ± 0.0332	724138.7000 ± 3275.1634	0.0%	0.0087 ± 0.0035	2703.4000 ± 1145.6095	nan ± nan	nan ± nan	0.0%	70/95
Q-learning ($\rho = 0.3, \epsilon = 0.5$)	2.12109 ± 0.0363	653307.1000 ± 3870.3261	0.0%	0.0074 ± 0.0042	2276.1000 ± 1112.2030	1.7295 ± 0.1521	535500.0000 ± 45500.0000	20.0%	67/145
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	2.41190 ± 0.0396	745194.1000 ± 7032.5614	0.0%	0.0121 ± 0.0074	3778.5000 ± 2284.2617	nan ± nan	nan ± nan	0.0%	90/179
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	2.88021 ± 0.0358	891604.8000 ± 5670.9228	0.0%	0.0088 ± 0.0041	2726.5000 ± 1287.0387	nan ± nan	nan ± nan	0.0%	63/294
Q-learning ($\rho = 0.05, \epsilon = 0.75$)	3.90533 ± 0.0632	1251083.3000 ± 10681.0254	0.0%	0.0104 ± 0.0043	3432.0000 ± 1388.5551	nan ± nan	nan ± nan	0.0%	45/111
Q-learning ($\rho = 0.1, \epsilon = 0.75$)	3.54679 ± 0.0558	1130245.9000 ± 8338.9072	0.0%	0.0100 ± 0.0039	3083.9000 ± 1364.1802	nan ± nan	nan ± nan	0.0%	57/100
Q-learning ($\rho = 0.3, \epsilon = 0.75$)	3.72662 ± 0.0574	1190927.0000 ± 11004.0365	0.0%	0.0040 ± 0.0018	1334.1000 ± 510.5324	nan ± nan	nan ± nan	0.0%	66/114
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	4.35389 ± 0.0683	1382378.1000 ± 15613.9938	0.0%	0.0104 ± 0.0078	3269.2000 ± 2697.0913	nan ± nan	nan ± nan	0.0%	77/168
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	5.15801 ± 0.0490	1636216.2000 ± 12417.5454	0.0%	0.0078 ± 0.0041	2408.0000 ± 1404.2232	nan ± nan	nan ± nan	0.0%	114/338
Q-learning ($\rho = 0.05, \epsilon = 0.9$)	6.85552 ± 0.0787	2236010.5000 ± 23359.9229	0.0%	0.0089 ± 0.0075	2856.5000 ± 2257.2585	nan ± nan	nan ± nan	0.0%	54/92
Q-learning ($\rho = 0.1, \epsilon = 0.9$)	6.82101 ± 0.0885	2225539.1000 ± 24245.5087	0.0%	0.0115 ± 0.0050	3586.1000 ± 1739.9655	nan ± nan	nan ± nan	0.0%	56/102
Q-learning ($\rho = 0.3, \epsilon = 0.9$)	7.53324 ± 0.1202	2449052.9000 ± 21388.0056	0.0%	0.0083 ± 0.0047	2627.3000 ± 1534.0655	nan ± nan	nan ± nan	0.0%	69/141
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	8.46074 ± 0.0970	2755312.1000 ± 31461.1213	0.0%	0.0091 ± 0.0047	2996.4000 ± 1609.3150	nan ± nan	nan ± nan	0.0%	66/297
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	9.42652 ± 0.1569	3065124.9000 ± 31097.3818	0.0%	0.0086 ± 0.0061	2463.8000 ± 1945.0780	nan ± nan	nan ± nan	0.0%	108/322
Q-learning ($\rho = 0.05, \epsilon = 1$)	15.74128 ± 0.1625	5184666.0000 ± 25176.4054	0.0%	0.0072 ± 0.0036	2279.2000 ± 1360.7437	nan ± nan	nan ± nan	0.0%	47/103
Q-learning ($\rho = 0.1, \epsilon = 1$)	15.82224 ± 0.2927	5225182.8000 ± 62938.6030	0.0%	0.0044 ± 0.0032	1509.3000 ± 1114.3018	nan ± nan	nan ± nan	0.0%	35/96
Q-learning ($\rho = 0.3, \epsilon = 1$)	17.44275 ± 1.1025	5251401.0000 ± 34692.6278	0.0%	0.0086 ± 0.0068	2264.4000 ± 1591.9069	nan ± nan	nan ± nan	0.0%	49/109
Q-learning ($\rho = 0.5, \epsilon = 1$)	18.25056 ± 0.5791	5224070.7000 ± 39479.1897	0.0%	0.0056 ± 0.0032	1583.6000 ± 827.3408	nan ± nan	nan ± nan	0.0%	56/175
Q-learning ($\rho = 0.7, \epsilon = 1$)	18.02579 ± 0.3255	5218716.7000 ± 57834.4366	0.0%	0.0095 ± 0.0096	2808.9000 ± 2980.6249	nan ± nan	nan ± nan	0.0%	99/254
Async Value Iteration	0.91119 ± 0.0214	1111232.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	49/100
Value Iteration	1.35879 ± 0.0292	1661120.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	52/94

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

Algorithm (Problem 1) ($\gamma = 0.7$)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.05, \epsilon = 0$)	1.93393 ± 0.0889	518018.8000 ± 1079.4022	0.0%	0.0062 ± 0.0027	1755.0000 ± 772.9291	nan ± nan	nan ± nan	0.0%	37/312
Q-learning ($\rho = 0.1, \epsilon = 0$)	1.22601 ± 0.0191	346009.9000 ± 1428.1871	0.0%	0.0086 ± 0.0030	2369.8000 ± 840.9631	1.2029 ± 0.0272	340500.0000 ± 4193.2485	60.0%	35/92
Q-learning ($\rho = 0.3, \epsilon = 0$)	0.75712 ± 0.0197	211921.6000 ± 1266.0164	0.0%	0.0059 ± 0.0026	1675.6000 ± 592.5429	0.4032 ± 0.0182	113700.0000 ± 3848.3763	100.0%	36/49
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.65779 ± 0.0115	186711.0000 ± 1534.7377	0.0%	0.0084 ± 0.0036	2143.8000 ± 1009.8904	0.2517 ± 0.0298	70900.0000 ± 8129.5756	100.0%	41/56
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.63893 ± 0.0130	183386.4000 ± 4281.1417	0.0%	0.0061 ± 0.0021	1684.2000 ± 597.2734	0.2239 ± 0.0784	64200.0000 ± 21469.9790	100.0%	35/70
Q-learning ($\rho = 0.05, \epsilon = 0.25$)	1.84066 ± 0.0490	532172.3000 ± 1821.7158	0.0%	0.0074 ± 0.0034	2220.0000 ± 1046.5392	nan ± nan	nan ± nan	0.0%	41/312
Q-learning ($\rho = 0.1, \epsilon = 0.25$)	1.28438 ± 0.0172	371955.9000 ± 800.9085	0.0%	0.0082 ± 0.0060	2265.8000 ± 1483.7717	1.2825 ± 0.0000	371000.0000 ± 0.0000	10.0%	34/55
Q-learning ($\rho = 0.3, \epsilon = 0.25$)	0.88986 ± 0.0169	258582.0000 ± 1501.4790	0.0%	0.0075 ± 0.0019	2033.1000 ± 549.4465	0.5190 ± 0.0621	151000.0000 ± 19271.7410	100.0%	32/98
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.86985 ± 0.0169	254281.8000 ± 690.4668	0.0%	0.0082 ± 0.0030	2355.4000 ± 835.7866	0.3277 ± 0.0637	96200.0000 ± 18258.1489	100.0%	36/73
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.96430 ± 0.0235	279320.7000 ± 2157.6759	0.0%	0.0076 ± 0.0025	2163.3000 ± 735.2480	0.2296 ± 0.0497	66000.0000 ± 14696.9385	100.0%	39/111
Q-learning ($\rho = 0.05, \epsilon = 0.5$)	1.95859 ± 0.0273	586129.8000 ± 2392.9607	0.0%	0.0074 ± 0.0028	2343.3000 ± 827.6835	nan ± nan	nan ± nan	0.0%	35/70
Q-learning ($\rho = 0.1, \epsilon = 0.5$)	1.50248 ± 0.0178	447338.7000 ± 2670.3218	0.0%	0.0096 ± 0.0037	2991.9000 ± 1129.5690	1.3737 ± 0.0000	412000.0000 ± 0.0000	10.0%	33/48
Q-learning ($\rho = 0.3, \epsilon = 0.5$)	1.20397 ± 0.0226	358492.8000 ± 1388.4665	0.0%	0.0086 ± 0.0056	2532.9000 ± 1536.8264	0.5031 ± 0.0000	155000.0000 ± 0.0000	10.0%	36/51
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	1.23370 ± 0.0241	366895.6000 ± 2016.7670	0.0%	0.0084 ± 0.0029	2662.9000 ± 776.2440	0.7441 ± 0.2714	223333.3333 ± 80400.3870	30.0%	39/135
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	1.36602 ± 0.0202	405297.1000 ± 2158.8903	0.0%	0.0095 ± 0.0041	2778.8000 ± 1035.1374	0.7672 ± 0.4735	226000.0000 ± 135000.0000	20.0%	49/206
Q-learning ($\rho = 0.05, \epsilon = 0.75$)	2.63461 ± 0.0437	804658.6000 ± 4958.7344	0.0%	0.0077 ± 0.0039	2453.1000 ± 1264.9200	nan ± nan	nan ± nan	0.0%	34/66
Q-learning ($\rho = 0.1, \epsilon = 0.75$)	2.23579 ± 0.0544	714200.6000 ± 4243.7842	0.0%	0.0083 ± 0.0040	2721.3000 ± 1333.0644	nan ± nan	nan ± nan	0.0%	35/54
Q-learning ($\rho = 0.3, \epsilon = 0.75$)	2.09277 ± 0.0412	671241.6000 ± 4798.8179	0.0%	0.0104 ± 0.0041	3375.1000 ± 1409.1686	nan ± nan	nan ± nan	0.0%	37/91
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	2.20245 ± 0.0270	702471.4000 ± 6924.6805	0.0%	0.0101 ± 0.0052	3063.8000 ± 1475.5833	nan ± nan	nan ± nan	0.0%	41/90
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	2.40047 ± 0.0310	778859.2000 ± 5927.0231	0.0%	0.0087 ± 0.0032	2838.8000 ± 1014.1625	nan ± nan	nan ± nan	0.0%	48/216
Q-learning ($\rho = 0.05, \epsilon = 0.9$)	4.63566 ± 0.0648	1529885.7000 ± 14965.9439	0.0%	0.0093 ± 0.0054	2950.3000 ± 1811.1419	nan ± nan	nan ± nan	0.0%	38/59
Q-learning ($\rho = 0.1, \epsilon = 0.9$)	4.53196 ± 0.0503	1504314.0000 ± 17157.8236	0.0%	0.0084 ± 0.0054	2891.9000 ± 1814.5724	nan ± nan	nan ± nan	0.0%	32/49
Q-learning ($\rho = 0.3, \epsilon = 0.9$)	4.69049 ± 0.1418	1518928.1000 ± 16357.8860	0.0%	0.0100 ± 0.0074	3207.4000 ± 2346.9705	nan ± nan	nan ± nan	0.0%	36/54
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	5.18126 ± 0.2639	1606007.0000 ± 14323.3288	0.0%	0.0106 ± 0.0074	2953.4000 ± 1984.1909	nan ± nan	nan ± nan	0.0%	39/67
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	5.32480 ± 0.0771	1747653.4000 ± 18328.3763	0.0%	0.0081 ± 0.0044	2636.9000 ± 1530.3755	nan ± nan	nan ± nan	0.0%	43/213
Q-learning ($\rho = 0.05, \epsilon = 1$)	14.37472 ± 0.1425	4828867.8000 ± 36170.1378	0.0%	0.0061 ± 0.0060	1831.1000 ± 2147.7063	nan ± nan	nan ± nan	0.0%	34/51
Q-learning ($\rho = 0.1, \epsilon = 1$)	14.45536 ± 0.2361	4827837.3000 ± 54592.7340	0.0%	0.0062 ± 0.0037	1993.1000 ± 1234.6338	nan ± nan	nan ± nan	0.0%	34/48
Q-learning ($\rho = 0.3, \epsilon = 1$)	14.30180 ± 0.1002	4800939.1000 ± 32144.3814	0.0%	0.0041 ± 0.0041	1327.7000 ± 1238.6299	nan ± nan	nan ± nan	0.0%	36/53
Q-learning ($\rho = 0.5, \epsilon = 1$)	14.39015 ± 0.1338	4811996.4000 ± 32298.9266	0.0%	0.0059 ± 0.0069	1954.0000 ± 2181.1991	nan ± nan	nan ± nan	0.0%	34/76
Q-learning ($\rho = 0.7, \epsilon = 1$)	14.34809 ± 0.1401	4807284.1000 ± 33911.3460	0.0%	0.0053 ± 0.0033	1504.8000 ± 1073.6604	nan ± nan	nan ± nan	0.0%	35/90
Async Value Iteration	0.33047 ± 0.0064	475424.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	38/50
Value Iteration	0.50604 ± 0.0121	721728.0000 ± 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 ρ in stochastic problem with $\gamma = 0.9$ (Problem 1).

Algorithm (Problem 1) ($\gamma = 0.9$)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.05, \epsilon = 0$)	1.40630 ± 0.0248	374388.4000 ± 465.3979	0.0%	0.0076 ± 0.0019	1905.9000 ± 403.3647	nan ± nan	nan ± nan	0.0%	36/1538
Q-learning ($\rho = 0.1, \epsilon = 0$)	0.94216 ± 0.0373	245359.1000 ± 529.2448	0.0%	0.0058 ± 0.0020	1530.3000 ± 411.9168	0.9268 ± 0.0408	241000.0000 ± 2236.0680	60.0%	29/149
Q-learning ($\rho = 0.3, \epsilon = 0$)	0.57358 ± 0.0309	146838.2000 ± 505.2702	0.0%	0.0052 ± 0.0013	1359.0000 ± 472.7898	0.3277 ± 0.0246	83100.0000 ± 3419.0642	100.0%	28/46
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.48362 ± 0.0215	128080.1000 ± 676.8250	0.0%	0.0054 ± 0.0015	1342.8000 ± 305.3411	0.1893 ± 0.0185	48900.0000 ± 1700.0000	100.0%	28/40
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.46465 ± 0.0163	122683.3000 ± 2014.7824	0.0%	0.0043 ± 0.0028	1179.8000 ± 420.3148	0.1604 ± 0.0362	41800.0000 ± 7166.5891	100.0%	28/37
Q-learning ($\rho = 0.05, \epsilon = 0.25$)	1.43774 ± 0.0240	387669.7000 ± 747.8249	0.0%	0.0066 ± 0.0028	1886.5000 ± 696.6657	nan ± nan	nan ± nan	0.0%	30/292
Q-learning ($\rho = 0.1, \epsilon = 0.25$)	0.97812 ± 0.0325	266944.7000 ± 588.4646	0.0%	0.0082 ± 0.0017	2255.3000 ± 486.6432	0.9239 ± 0.0000	263000.0000 ± 0.0000	10.0%	28/34
Q-learning ($\rho = 0.3, \epsilon = 0.25$)	0.65382 ± 0.0262	178714.2000 ± 602.5748	0.0%	0.0062 ± 0.0009	1630.3000 ± 348.9653	0.5085 ± 0.0471	136000.0000 ± 10977.2492	40.0%	30/39
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.60797 ± 0.0249	162876.3000 ± 617.6632	0.0%	0.0058 ± 0.0018	1642.6000 ± 419.5217	0.3219 ± 0.1023	85777.7778 ± 24530.1528	90.0%	30/43
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.59050 ± 0.0233	162027.3000 ± 757.8191	0.0%	0.0067 ± 0.0022	1651.8000 ± 461.5987	0.2097 ± 0.0428	58200.0000 ± 10097.5244	100.0%	28/87
Q-learning ($\rho = 0.05, \epsilon = 0.5$)	1.53399 ± 0.0232	429355.3000 ± 1100.5893	0.0%	0.0073 ± 0.0026	2106.6000 ± 748.5252	nan ± nan	nan ± nan	0.0%	30/284
Q-learning ($\rho = 0.1, \epsilon = 0.5$)	1.14633 ± 0.0310	324285.5000 ± 1215.7992	0.0%	0.0080 ± 0.0027	2126.5000 ± 517.5400	1.0349 ± 0.0000	295000.0000 ± 0.0000	10.0%	29/38
Q-learning ($\rho = 0.3, \epsilon = 0.5$)	0.90733 ± 0.0418	251963.2000 ± 1651.1753	0.0%	0.0074 ± 0.0029	2008.2000 ± 685.7393	nan ± nan	nan ± nan	0.0%	28/35
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.83558 ± 0.0203	237662.9000 ± 1018.6010	0.0%	0.0068 ± 0.0016	1893.6000 ± 501.3544	0.8210 ± 0.0000	236000.0000 ± 0.0000	10.0%	28/34
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.84558 ± 0.0386	238293.1000 ± 1387.4264	0.0%	0.0070 ± 0.0017	1948.5000 ± 425.8716	0.3798 ± 0.1223	107250.0000 ± 33461.7319	40.0%	29/94
Q-learning ($\rho = 0.05, \epsilon = 0.75$)	2.06221 ± 0.0449	596335.3000 ± 2529.3901	0.0%	0.0098 ± 0.0042	2738.5000 ± 1076.0376	nan ± nan	nan ± nan	0.0%	30/40
Q-learning ($\rho = 0.1, \epsilon = 0.75$)	1.83900 ± 0.0367	526975.9000 ± 2097.9857	0.0%	0.0084 ± 0.0026	2374.0000 ± 784.2525	nan ± nan	nan ± nan	0.0%	29/37
Q-learning ($\rho = 0.3, \epsilon = 0.75$)	1.66193 ± 0.0250	479980.7000 ± 3110.5655	0.0%	0.0087 ± 0.0031	2513.5000 ± 931.2594	nan ± nan	nan ± nan	0.0%	28/34
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	1.63547 ± 0.0554	473658.0000 ± 2320.7265	0.0%	0.0100 ± 0.0043	2728.8000 ± 1042.9370	0.5347 ± 0.0000	155000.0000 ± 0.0000	10.0%	29/38
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	1.67548 ± 0.0434	478412.8000 ± 2841.9262	0.0%	0.0086 ± 0.0022	2461.2000 ± 788.4459	nan ± nan	nan ± nan	0.0%	28/87
Q-learning ($\rho = 0.05, \epsilon = 0.9$)	3.89552 ± 0.0531	1165573.5000 ± 10356.0597	0.0%	0.0104 ± 0.0072	3129.3000 ± 2040.2814	nan ± nan	nan ± nan	0.0%	29/35
Q-learning ($\rho = 0.1, \epsilon = 0.9$)	3.83259 ± 0.0690	1136885.5000 ± 5079.6989	0.0%	0.0098 ± 0.0053	2640.0000 ± 1763.6383	nan ± nan	nan ± nan	0.0%	29/36
Q-learning ($\rho = 0.3, \epsilon = 0.9$)	3.87358 ± 0.0763	1133814.1000 ± 8302.0288	0.0%	0.0074 ± 0.0042	2248.7000 ± 1257.7827	nan ± nan	nan ± nan	0.0%	29/36
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	3.91092 ± 0.0531	1141361.2000 ± 11669.4104	0.0%	0.0109 ± 0.0076	2984.7000 ± 2097.6476	nan ± nan	nan ± nan	0.0%	29/37
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	3.93309 ± 0.0699	1154820.9000 ± 7291.7631	0.0%	0.0096 ± 0.0043	2808.7000 ± 1300.1570	nan ± nan	nan ± nan	0.0%	29/147
Q-learning ($\rho = 0.05, \epsilon = 1$)	15.01271 ± 0.1990	4469635.5000 ± 51363.0854	0.0%	0.0078 ± 0.0071	2407.0000 ± 2354.6838	nan ± nan	nan ± nan	0.0%	28/34
Q-learning ($\rho = 0.1, \epsilon = 1$)	14.97888 ± 0.1849	4478636.0000 ± 62462.4331	0.0%	0.0058 ± 0.0034	1641.1000 ± 899.9632	nan ± nan	nan ± nan	0.0%	30/37
Q-learning ($\rho = 0.3, \epsilon = 1$)	14.73030 ± 0.2281	4472910.4000 ± 42569.9358	0.0%	0.0065 ± 0.0048	1889.7000 ± 1192.4006	nan ± nan	nan ± nan	0.0%	29/38
Q-learning ($\rho = 0.5, \epsilon = 1$)	15.22034 ± 0.3940	4481033.5000 ± 43593.8014	0.0%	0.0065 ± 0.0082	1784.5000 ± 2133.5604	nan ± nan	nan ± nan	0.0%	29/35
Q-learning ($\rho = 0.7, \epsilon = 1$)	15.08263 ± 0.3920	4457994.9000 ± 22794.1670	0.0%	0.0035 ± 0.0017	981.9000 ± 535.6134	nan ± nan	nan ± nan	0.0%	32/75
Async Value Iteration	0.18373 ± 0.0165	226256.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	28/34
Value Iteration	0.27155 ± 0.0040	343680.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	29/36

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

Algorithm (Problem 10) ($\gamma = 0.5$)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.05, \epsilon = 0$)	9.40443 ± 0.1140	2480472.9000 ± 7745.3471	0.0%	0.1200 ± 0.0947	32601.0000 ± 25738.4943	nan ± nan	nan ± nan	0.0%	238/445
Q-learning ($\rho = 0.1, \epsilon = 0$)	6.02799 ± 0.1197	1620228.2000 ± 7376.2157	0.0%	0.0939 ± 0.0348	25728.8000 ± 9519.7926	5.7811 ± 0.1068	1553000.0000 ± 18606.4505	100.0%	163/232
Q-learning ($\rho = 0.3, \epsilon = 0$)	3.54922 ± 0.0715	975451.9000 ± 9655.4821	0.0%	0.0452 ± 0.0357	12364.1000 ± 9482.9381	1.9270 ± 0.1422	532714.2857 ± 43839.6894	70.0%	142/263
Q-learning ($\rho = 0.5, \epsilon = 0$)	3.17621 ± 0.0505	879102.4000 ± 13959.9244	0.0%	0.0405 ± 0.0282	11513.0000 ± 7962.2145	1.6239 ± 0.6521	447666.6667 ± 178449.4949	30.0%	197/253
Q-learning ($\rho = 0.7, \epsilon = 0$)	3.12813 ± 0.1039	865605.3000 ± 28493.4429	0.0%	0.0335 ± 0.0212	9296.6000 ± 6125.0827	1.0968 ± 0.3691	304000.0000 ± 92245.3251	50.0%	175/286
Q-learning ($\rho = 0.05, \epsilon = 0.25$)	9.12259 ± 0.0575	2554175.1000 ± 11350.4274	0.0%	0.0876 ± 0.0438	24976.2000 ± 13129.7253	nan ± nan	nan ± nan	0.0%	182/493
Q-learning ($\rho = 0.1, \epsilon = 0.25$)	6.26658 ± 0.0751	1754951.5000 ± 5525.9510	0.0%	0.0659 ± 0.0203	18746.5000 ± 5830.1253	6.0391 ± 0.1214	1689000.0000 ± 20542.6386	90.0%	133/347
Q-learning ($\rho = 0.3, \epsilon = 0.25$)	4.68124 ± 0.0436	1311938.9000 ± 6453.3623	0.0%	0.0395 ± 0.0270	11493.1000 ± 7651.9357	2.0760 ± 0.0508	588000.0000 ± 9736.5292	50.0%	139/384
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	5.16458 ± 0.1555	1417405.1000 ± 7168.2131	0.0%	0.0511 ± 0.0269	14874.7000 ± 7885.9615	1.3483 ± 0.0346	370285.7143 ± 7458.7300	70.0%	225/411
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	6.42202 ± 0.0553	1705747.2000 ± 6643.0314	0.0%	0.0512 ± 0.0240	13308.6000 ± 6114.1130	1.0720 ± 0.0232	285333.3333 ± 2867.4418	30.0%	300/1142
Q-learning ($\rho = 0.05, \epsilon = 0.5$)	9.91728 ± 0.2256	2775826.5000 ± 9967.5108	0.0%	0.1145 ± 0.0646	32119.6000 ± 17678.3821	nan ± nan	nan ± nan	0.0%	178/292
Q-learning ($\rho = 0.1, \epsilon = 0.5$)	7.36858 ± 0.1876	2075725.3000 ± 11153.2485	0.0%	0.0898 ± 0.0339	25407.8000 ± 9106.8728	7.0158 ± 0.2400	1948833.3333 ± 32789.3106	60.0%	137/213
Q-learning ($\rho = 0.3, \epsilon = 0.5$)	6.49163 ± 0.0532	1823326.1000 ± 7409.6966	0.0%	0.0872 ± 0.0454	24703.0000 ± 12189.4823	2.5204 ± 0.0687	707500.0000 ± 20377.6839	100.0%	156/267
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	7.45448 ± 0.0760	2091213.2000 ± 10470.8515	0.0%	0.0646 ± 0.0176	18269.6000 ± 4873.1855	2.8333 ± 1.1576	794800.0000 ± 319459.7940	100.0%	234/535
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	8.98784 ± 0.0527	2538064.6000 ± 10811.5659	0.0%	0.0493 ± 0.0321	13888.5000 ± 9257.8329	nan ± nan	nan ± nan	0.0%	207/837
Q-learning ($\rho = 0.05, \epsilon = 0.75$)	12.58560 ± 0.1962	3677258.2000 ± 19942.8683	0.0%	0.1077 ± 0.1077	32086.5000 ± 31816.4864	nan ± nan	nan ± nan	0.0%	138/221
Q-learning ($\rho = 0.1, \epsilon = 0.75$)	10.76710 ± 0.0524	3205292.3000 ± 13684.9267	0.0%	0.1102 ± 0.0643	33290.5000 ± 19055.9031	8.6584 ± 0.0000	2599000.0000 ± 0.0000	10.0%	163/221
Q-learning ($\rho = 0.3, \epsilon = 0.75$)	11.34678 ± 0.0784	3362790.6000 ± 17116.1092	0.0%	0.0491 ± 0.0325	14922.7000 ± 10020.2733	nan ± nan	nan ± nan	0.0%	175/298
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	13.50038 ± 0.1144	3983012.0000 ± 27323.2308	0.0%	0.0730 ± 0.0457	21650.6000 ± 13761.3699	nan ± nan	nan ± nan	0.0%	207/632
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	16.31277 ± 0.1463	4831696.4000 ± 20442.7072	0.0%	0.0582 ± 0.0399	16911.0000 ± 10777.8656	nan ± nan	nan ± nan	0.0%	284/854
Q-learning ($\rho = 0.05, \epsilon = 0.9$)	20.90885 ± 0.5429	6336251.9000 ± 24239.5722	0.0%	0.1113 ± 0.0526	33965.9000 ± 16003.0625	nan ± nan	nan ± nan	0.0%	147/235
Q-learning ($\rho = 0.1, \epsilon = 0.9$)	20.46911 ± 0.3372	6326221.1000 ± 41608.1798	0.0%	0.1141 ± 0.0729	34939.6000 ± 22734.0983	nan ± nan	nan ± nan	0.0%	159/209
Q-learning ($\rho = 0.3, \epsilon = 0.9$)	21.71623 ± 0.3406	6846501.8000 ± 40375.1530	0.0%	0.0475 ± 0.0496	14980.1000 ± 15556.5282	nan ± nan	nan ± nan	0.0%	140/434
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	23.11406 ± 0.6263	7398753.6000 ± 41900.8090	0.0%	0.1019 ± 0.0624	33511.3000 ± 20074.2052	nan ± nan	nan ± nan	0.0%	207/562
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	24.36300 ± 0.5107	7848426.5000 ± 25471.5155	0.0%	0.1275 ± 0.0920	41382.4000 ± 28358.8550	nan ± nan	nan ± nan	0.0%	226/724
Q-learning ($\rho = 0.05, \epsilon = 1$)	26.72804 ± 0.4123	8865132.9000 ± 9711.4018	0.0%	0.1990 ± 0.1340	66093.1000 ± 45004.6923	nan ± nan	nan ± nan	0.0%	123/263
Q-learning ($\rho = 0.1, \epsilon = 1$)	69.30524 ± 126.7632	8866112.7000 ± 9800.4870	0.0%	0.2255 ± 0.1916	73164.3000 ± 61230.6205	nan ± nan	nan ± nan	0.0%	136/242
Q-learning ($\rho = 0.3, \epsilon = 1$)	25.72737 ± 0.0487	8871062.7000 ± 6712.5570	0.0%	0.2037 ± 0.1301	69453.3000 ± 42509.9341	nan ± nan	nan ± nan	0.0%	182/299
Q-learning ($\rho = 0.5, \epsilon = 1$)	25.71404 ± 0.0721	8871489.3000 ± 10760.8833	0.0%	0.2117 ± 0.2134	74120.4000 ± 73008.9996	nan ± nan	nan ± nan	0.0%	204/501
Q-learning ($\rho = 0.7, \epsilon = 1$)	25.77060 ± 0.1262	8866343.3000 ± 9861.5639	0.0%	0.1483 ± 0.1428	50913.4000 ± 49458.5697	nan ± nan	nan ± nan	0.0%	249/568
Async Value Iteration	1.09475 ± 0.0118	1564666.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	146/191
Value Iteration	1.68654 ± 0.0290	2360832.0000 ± 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 ρ in stochastic problem with $\gamma = 0.7$ (Problem 10).

Algorithm (Problem 10) ($\gamma = 0.7$)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.05, \epsilon = 0$)	5.96387 ± 0.1696	1512161.4000 ± 2670.5442	0.0%	0.0579 ± 0.0476	15139.5000 ± 12453.2009	nan ± nan	nan ± nan	0.0%	119/590
Q-learning ($\rho = 0.1, \epsilon = 0$)	3.78183 ± 0.0686	954735.8000 ± 2747.9469	0.0%	0.0444 ± 0.0276	11433.3000 ± 7109.2537	3.6423 ± 0.0671	923375.0000 ± 15826.6982	80.0%	89/126
Q-learning ($\rho = 0.3, \epsilon = 0$)	2.16876 ± 0.0554	552439.9000 ± 7708.0647	0.0%	0.0339 ± 0.0230	8278.7000 ± 5430.2361	1.1778 ± 0.0485	299500.0000 ± 3354.1020	100.0%	103/156
Q-learning ($\rho = 0.5, \epsilon = 0$)	1.91131 ± 0.0531	486992.3000 ± 9675.5909	0.0%	0.0178 ± 0.0115	4832.2000 ± 3207.7101	0.9841 ± 0.2493	250666.6667 ± 63712.8977	90.0%	93/135
Q-learning ($\rho = 0.7, \epsilon = 0$)	1.88206 ± 0.0723	479389.7000 ± 11571.0215	0.0%	0.0170 ± 0.0097	4054.9000 ± 2071.4557	0.8854 ± 0.4787	224555.5556 ± 118481.2326	90.0%	102/164
Q-learning ($\rho = 0.05, \epsilon = 0.25$)	6.02142 ± 0.0394	1562032.3000 ± 2398.1777	0.0%	0.0386 ± 0.0308	9979.9000 ± 7808.7939	nan ± nan	nan ± nan	0.0%	104/722
Q-learning ($\rho = 0.1, \epsilon = 0.25$)	4.00974 ± 0.0267	1035449.3000 ± 3159.8627	0.0%	0.0272 ± 0.0302	6948.8000 ± 7603.1372	3.8178 ± 0.0049	987000.0000 ± 1000.0000	20.0%	95/125
Q-learning ($\rho = 0.3, \epsilon = 0.25$)	2.66555 ± 0.0287	692706.6000 ± 3277.2031	0.0%	0.0261 ± 0.0161	7338.9000 ± 4366.2303	1.3116 ± 0.0418	340500.0000 ± 11800.4237	100.0%	104/128
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	2.65333 ± 0.0295	689653.7000 ± 3702.4202	0.0%	0.0254 ± 0.0178	6484.2000 ± 4885.6365	0.8223 ± 0.0439	213100.0000 ± 11835.9621	100.0%	97/197
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	2.95973 ± 0.0557	771130.4000 ± 3641.6586	0.0%	0.0272 ± 0.0171	7063.4000 ± 4201.0633	0.6182 ± 0.0096	155666.6667 ± 1490.7120	60.0%	148/296
Q-learning ($\rho = 0.05, \epsilon = 0.5$)	6.51836 ± 0.0778	1718152.0000 ± 4763.9567	0.0%	0.0763 ± 0.0425	20883.3000 ± 11174.4661	nan ± nan	nan ± nan	0.0%	92/184
Q-learning ($\rho = 0.1, \epsilon = 0.5$)	4.91298 ± 0.0634	1245892.1000 ± 5218.7915	0.0%	0.0364 ± 0.0220	9807.3000 ± 5512.0979	4.6469 ± 0.0429	1190250.0000 ± 10825.3175	40.0%	85/127
Q-learning ($\rho = 0.3, \epsilon = 0.5$)	3.74506 ± 0.0729	965362.3000 ± 2055.3866	0.0%	0.0380 ± 0.0186	10192.3000 ± 4855.7640	1.7846 ± 0.2619	460200.0000 ± 65978.4813	100.0%	100/142
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	3.86460 ± 0.0834	1001219.4000 ± 2588.9391	0.0%	0.0234 ± 0.0165	6067.5000 ± 4365.0193	0.9733 ± 0.0582	255600.0000 ± 18413.0389	100.0%	99/179
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	4.36818 ± 0.0934	1129015.9000 ± 5753.2040	0.0%	0.0216 ± 0.0094	5799.7000 ± 2554.9131	0.7853 ± 0.0964	205800.0000 ± 25646.8322	100.0%	143/428
Q-learning ($\rho = 0.05, \epsilon = 0.75$)	8.75981 ± 0.1488	2329030.8000 ± 5585.0317	0.0%	0.1138 ± 0.0793	31149.9000 ± 22031.2859	nan ± nan	nan ± nan	0.0%	79/119
Q-learning ($\rho = 0.1, \epsilon = 0.75$)	7.43980 ± 0.1085	1989297.4000 ± 7094.2379	0.0%	0.0618 ± 0.0334	16474.3000 ± 8929.8288	5.7761 ± 0.5224	1556250.0000 ± 133329.2447	40.0%	92/147
Q-learning ($\rho = 0.3, \epsilon = 0.75$)	6.88314 ± 0.1305	1831675.0000 ± 6392.2100	0.0%	0.0517 ± 0.0259	13712.3000 ± 7438.6561	3.8212 ± 1.7151	1024000.0000 ± 465013.1181	50.0%	101/121
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	7.16563 ± 0.1659	1942509.1000 ± 8138.6615	0.0%	0.0394 ± 0.0226	10599.0000 ± 5948.5857	nan ± nan	nan ± nan	0.0%	118/197
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	8.00057 ± 0.0740	2178969.9000 ± 6554.5616	0.0%	0.0390 ± 0.0243	10646.1000 ± 6928.9696	nan ± nan	nan ± nan	0.0%	144/436
Q-learning ($\rho = 0.05, \epsilon = 0.9$)	15.95728 ± 0.1505	4445227.4000 ± 29971.9655	0.0%	0.0858 ± 0.0549	24813.2000 ± 16237.4335	11.2460 ± 0.0000	3143000.0000 ± 0.0000	10.0%	84/143
Q-learning ($\rho = 0.1, \epsilon = 0.9$)	15.71833 ± 0.2596	4333234.7000 ± 28248.6697	0.0%	0.0887 ± 0.0443	24420.4000 ± 12540.7841	nan ± nan	nan ± nan	0.0%	97/130
Q-learning ($\rho = 0.3, \epsilon = 0.9$)	16.10899 ± 0.2476	4420356.4000 ± 13032.1730	0.0%	0.0728 ± 0.0468	19668.4000 ± 11879.1242	nan ± nan	nan ± nan	0.0%	90/198
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	17.26956 ± 0.2914	4729963.8000 ± 24146.7802	0.0%	0.0708 ± 0.0508	19886.5000 ± 14764.4005	nan ± nan	nan ± nan	0.0%	96/160
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	19.28740 ± 0.3056	5192691.0000 ± 29151.5206	0.0%	0.0596 ± 0.0344	15630.2000 ± 9230.4066	nan ± nan	nan ± nan	0.0%	131/321
Q-learning ($\rho = 0.05, \epsilon = 1$)	31.51687 ± 0.8513	8822073.6000 ± 19012.0775	0.0%	0.1775 ± 0.1251	52010.1000 ± 36275.3903	nan ± nan	nan ± nan	0.0%	88/132
Q-learning ($\rho = 0.1, \epsilon = 1$)	29.13096 ± 0.3870	8812052.7000 ± 13673.2322	0.0%	0.0965 ± 0.1010	29015.0000 ± 30287.0171	nan ± nan	nan ± nan	0.0%	86/126
Q-learning ($\rho = 0.3, \epsilon = 1$)	28.76155 ± 0.3539	8808772.4000 ± 9103.4225	0.0%	0.1630 ± 0.1231	51002.2000 ± 38101.9993	nan ± nan	nan ± nan	0.0%	82/128
Q-learning ($\rho = 0.5, \epsilon = 1$)	28.78621 ± 0.5184	8812472.3000 ± 13934.9893	0.0%	0.1213 ± 0.0657	38426.4000 ± 20768.8467	nan ± nan	nan ± nan	0.0%	100/268
Q-learning ($\rho = 0.7, \epsilon = 1$)	28.99773 ± 0.3141	8805165.2000 ± 16054.1824	0.0%	0.2151 ± 0.2074	67427.7000 ± 64089.7856	nan ± nan	nan ± nan	0.0%	109/275
Async Value Iteration	0.55572 ± 0.0106	697798.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	89/124
Value Iteration	0.89072 ± 0.0274	1075900.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	82/117

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

Algorithm (Problem 10) ($\gamma = 0.9$)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.05, \epsilon = 0$)	4.25422 ± 0.1022	1074244.4000 ± 843.2783	0.0%	0.0286 ± 0.0176	7168.8000 ± 4539.2867	nan ± nan	nan ± nan	0.0%	84/29307
Q-learning ($\rho = 0.1, \epsilon = 0$)	2.47363 ± 0.0346	661308.6000 ± 562.9441	0.0%	0.0142 ± 0.0190	4155.8000 ± 5454.1404	2.4213 ± 0.0263	645666.6667 ± 2867.4418	60.0%	67/78
Q-learning ($\rho = 0.3, \epsilon = 0$)	1.38754 ± 0.0357	372630.7000 ± 1890.6645	0.0%	0.0098 ± 0.0083	2771.1000 ± 2129.7503	0.7891 ± 0.0331	213500.0000 ± 8958.2364	100.0%	69/84
Q-learning ($\rho = 0.5, \epsilon = 0$)	1.20380 ± 0.0440	321507.2000 ± 5933.9071	0.0%	0.0154 ± 0.0062	4374.7000 ± 1932.6496	0.5243 ± 0.0804	138200.0000 ± 17803.3705	100.0%	73/91
Q-learning ($\rho = 0.7, \epsilon = 0$)	1.15708 ± 0.0334	309478.0000 ± 5366.3737	0.0%	0.0062 ± 0.0025	1862.9000 ± 916.8312	0.5458 ± 0.3031	143000.0000 ± 75529.2438	60.0%	71/89
Q-learning ($\rho = 0.05, \epsilon = 0.25$)	4.09429 ± 0.0360	1118923.6000 ± 1164.5843	0.0%	0.0295 ± 0.0161	8258.0000 ± 4771.8050	nan ± nan	nan ± nan	0.0%	73/831
Q-learning ($\rho = 0.1, \epsilon = 0.25$)	2.60927 ± 0.0388	728539.7000 ± 866.2772	0.0%	0.0220 ± 0.0132	5969.8000 ± 3317.7490	2.5393 ± 0.0587	706800.0000 ± 18882.7964	50.0%	66/84
Q-learning ($\rho = 0.3, \epsilon = 0.25$)	1.62901 ± 0.0252	460307.9000 ± 870.4154	0.0%	0.0201 ± 0.0095	5577.4000 ± 2559.6853	1.0746 ± 0.0955	305857.1429 ± 28975.0068	70.0%	67/88
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	1.49651 ± 0.0315	418517.9000 ± 1674.5908	0.0%	0.0158 ± 0.0071	4227.6000 ± 1757.0224	0.5465 ± 0.0697	153400.0000 ± 20140.5064	100.0%	67/484
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	1.52844 ± 0.0172	432750.2000 ± 2054.6122	0.0%	0.0118 ± 0.0071	3287.1000 ± 2216.3088	0.4230 ± 0.0674	120500.0000 ± 20323.6316	100.0%	78/400
Q-learning ($\rho = 0.05, \epsilon = 0.5$)	4.28771 ± 0.0479	1236475.8000 ± 2899.7580	0.0%	0.0324 ± 0.0213	9622.1000 ± 6321.6981	nan ± nan	nan ± nan	0.0%	70/388
Q-learning ($\rho = 0.1, \epsilon = 0.5$)	3.07881 ± 0.0272	889892.2000 ± 1814.2900	0.0%	0.0281 ± 0.0191	8646.2000 ± 5451.9187	2.8183 ± 0.0983	813375.0000 ± 33540.7867	80.0%	66/86
Q-learning ($\rho = 0.3, \epsilon = 0.5$)	2.23190 ± 0.0341	658970.8000 ± 2453.3467	0.0%	0.0189 ± 0.0137	5419.8000 ± 3722.3724	1.1581 ± 0.2314	343000.0000 ± 69415.9307	70.0%	68/82
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	2.10181 ± 0.0232	622804.0000 ± 1562.8830	0.0%	0.0187 ± 0.0084	5537.9000 ± 2506.8575	0.6816 ± 0.1490	202400.0000 ± 43518.2720	100.0%	70/135
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	2.15394 ± 0.0292	636542.0000 ± 1429.2440	0.0%	0.0108 ± 0.0085	3264.2000 ± 2427.8488	0.4365 ± 0.0299	129000.0000 ± 8763.5609	100.0%	75/269
Q-learning ($\rho = 0.05, \epsilon = 0.75$)	5.65363 ± 0.0983	1703389.6000 ± 6419.9470	0.0%	0.0545 ± 0.0373	16982.8000 ± 11936.7565	nan ± nan	nan ± nan	0.0%	69/76
Q-learning ($\rho = 0.1, \epsilon = 0.75$)	4.78082 ± 0.0340	1446867.9000 ± 4909.8018	0.0%	0.0502 ± 0.0218	15544.9000 ± 6677.8183	3.5025 ± 0.5400	1062250.0000 ± 158316.7316	80.0%	65/77
Q-learning ($\rho = 0.3, \epsilon = 0.75$)	4.27364 ± 0.0704	1287621.8000 ± 5580.9577	0.0%	0.0283 ± 0.0179	8946.7000 ± 5479.3997	2.2220 ± 0.6976	666777.7778 ± 208290.8105	90.0%	69/83
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	4.17856 ± 0.0413	1270217.3000 ± 8410.6708	0.0%	0.0226 ± 0.0117	6767.0000 ± 3345.5285	1.3636 ± 0.5554	415000.0000 ± 171860.9903	100.0%	66/152
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	4.28299 ± 0.0702	1297171.0000 ± 7627.5490	0.0%	0.0280 ± 0.0137	8349.1000 ± 4065.5956	1.1803 ± 0.4989	357700.0000 ± 144956.5797	100.0%	76/138
Q-learning ($\rho = 0.05, \epsilon = 0.9$)	10.95564 ± 0.1662	3385292.5000 ± 21731.9319	0.0%	0.0418 ± 0.0216	13515.3000 ± 7494.2332	8.4737 ± 1.1036	2621111.1111 ± 328960.8029	90.0%	75/86
Q-learning ($\rho = 0.1, \epsilon = 0.9$)	10.54643 ± 0.0798	3277532.3000 ± 18161.0016	0.0%	0.0419 ± 0.0162	12685.3000 ± 4823.0110	5.4080 ± 1.4687	1682625.0000 ± 448392.3888	80.0%	66/79
Q-learning ($\rho = 0.3, \epsilon = 0.9$)	10.37067 ± 0.0760	3219723.5000 ± 18550.0356	0.0%	0.0292 ± 0.0231	9203.1000 ± 7032.5092	3.3454 ± 1.8469	1038285.7143 ± 569908.4423	70.0%	68/82
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	10.75918 ± 0.2029	3232790.3000 ± 21805.9856	0.0%	0.0315 ± 0.0183	9817.8000 ± 5485.1292	4.5028 ± 2.5028	1378833.3333 ± 796713.1263	60.0%	69/111
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	10.81855 ± 0.2442	3299951.7000 ± 19867.0822	0.0%	0.0513 ± 0.0290	15617.3000 ± 8658.5057	2.2075 ± 0.5192	685500.0000 ± 165500.0000	20.0%	71/213
Q-learning ($\rho = 0.05, \epsilon = 1$)	28.41448 ± 0.5752	8751202.5000 ± 9411.5658	0.0%	0.0975 ± 0.0623	30976.8000 ± 19431.4598	26.8243 ± 0.4043	8315000.0000 ± 268000.0000	20.0%	72/83
Q-learning ($\rho = 0.1, \epsilon = 1$)	28.09445 ± 0.3950	8746537.1000 ± 17063.6414	0.0%	0.1330 ± 0.1237	42667.7000 ± 39577.3723	21.1432 ± 4.3795	6613375.0000 ± 1345175.1687	80.0%	70/82
Q-learning ($\rho = 0.3, \epsilon = 1$)	28.15691 ± 0.4088	8735736.8000 ± 16185.1551	0.0%	0.1008 ± 0.0978	31559.3000 ± 30688.4362	12.3954 ± 5.6479	3848300.0000 ± 1720338.4580	100.0%	66/84
Q-learning ($\rho = 0.5, \epsilon = 1$)	30.57344 ± 0.6988	8749544.5000 ± 12582.7589	0.0%	0.0891 ± 0.1044	25791.4000 ± 31442.1964	7.8898 ± 4.4143	2273800.0000 ± 1290161.6023	100.0%	71/181
Q-learning ($\rho = 0.7, \epsilon = 1$)	30.92785 ± 0.3097	8748932.6000 ± 10139.2903	0.0%	0.0439 ± 0.0194	13493.7000 ± 6125.6220	16.7331 ± 9.8434	4786600.0000 ± 2805490.0891	50.0%	68/86
Async Value Iteration	0.29068 ± 0.0101	347362.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	67/87
Value Iteration	0.48309 ± 0.0289	553320.0000 ± 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 ρ in stochastic problem with $\gamma = 0.5$ (Problem 11).

Algorithm (Problem 11) ($\gamma = 0.5$)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.05, \epsilon = 0$)	3.05628 ± 0.0417	931903.1000 ± 4335.6100	0.0%	0.0105 ± 0.0047	3205.2000 ± 1516.4140	nan ± nan	nan ± nan	0.0%	38/238
Q-learning ($\rho = 0.1, \epsilon = 0$)	1.99090 ± 0.0282	608600.1000 ± 2705.2238	0.0%	0.0089 ± 0.0055	2631.0000 ± 1653.4711	1.7480 ± 0.0448	533700.0000 ± 8414.8678	100.0%	39/306
Q-learning ($\rho = 0.3, \epsilon = 0$)	1.19195 ± 0.0204	363139.1000 ± 8142.2723	0.0%	0.0080 ± 0.0030	2522.5000 ± 971.9032	0.6873 ± 0.1315	210000.0000 ± 40281.5094	100.0%	43/76
Q-learning ($\rho = 0.5, \epsilon = 0$)	1.05041 ± 0.0230	318803.0000 ± 6969.1476	0.0%	0.0075 ± 0.0036	2453.6000 ± 1079.3368	0.4852 ± 0.1418	148222.2222 ± 44681.0369	90.0%	47/109
Q-learning ($\rho = 0.7, \epsilon = 0$)	1.03261 ± 0.0488	316623.4000 ± 14922.6060	0.0%	0.0090 ± 0.0054	2880.2000 ± 1719.5827	0.3469 ± 0.1628	109125.0000 ± 51933.7017	80.0%	59/126
Q-learning ($\rho = 0.05, \epsilon = 0.25$)	3.09059 ± 0.0334	962644.0000 ± 4848.7360	0.0%	0.0114 ± 0.0049	3574.5000 ± 1554.0538	nan ± nan	nan ± nan	0.0%	40/258
Q-learning ($\rho = 0.1, \epsilon = 0.25$)	2.12340 ± 0.0382	660166.5000 ± 2695.4398	0.0%	0.0123 ± 0.0036	3998.8000 ± 1317.5341	1.8310 ± 0.0649	570000.0000 ± 14906.3745	100.0%	48/95
Q-learning ($\rho = 0.3, \epsilon = 0.25$)	1.54959 ± 0.0322	481027.4000 ± 3625.3863	0.0%	0.0108 ± 0.0063	3528.0000 ± 2148.9767	0.7006 ± 0.0974	219800.0000 ± 30632.6623	100.0%	38/88
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	1.67900 ± 0.0304	514450.0000 ± 4841.6683	0.0%	0.0083 ± 0.0029	2706.6000 ± 968.5790	0.7044 ± 0.4142	216571.4286 ± 122715.5330	70.0%	68/310
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	2.02010 ± 0.0134	618068.0000 ± 4699.3731	0.0%	0.0132 ± 0.0067	3988.8000 ± 2111.5785	0.3730 ± 0.1942	116500.0000 ± 60911.8215	100.0%	102/279
Q-learning ($\rho = 0.05, \epsilon = 0.5$)	3.31060 ± 0.0454	1051958.0000 ± 7819.2191	0.0%	0.0102 ± 0.0050	3123.4000 ± 1505.8675	nan ± nan	nan ± nan	0.0%	51/146
Q-learning ($\rho = 0.1, \epsilon = 0.5$)	2.47391 ± 0.0344	780987.2000 ± 5454.8317	0.0%	0.0098 ± 0.0069	3081.4000 ± 1978.9308	1.9975 ± 0.0654	631888.8889 ± 17291.2595	90.0%	35/77
Q-learning ($\rho = 0.3, \epsilon = 0.5$)	2.10076 ± 0.0147	667489.6000 ± 3971.0397	0.0%	0.0146 ± 0.0062	4812.3000 ± 1918.8502	0.6891 ± 0.0249	221800.0000 ± 7717.5126	100.0%	38/109
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	2.37681 ± 0.0257	759175.7000 ± 4896.8396	0.0%	0.0118 ± 0.0060	3729.3000 ± 1942.0390	0.5349 ± 0.1143	174300.0000 ± 38946.2450	100.0%	66/301
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	2.92918 ± 0.0494	926939.8000 ± 10533.7296	0.0%	0.0124 ± 0.0065	4034.1000 ± 2156.0668	1.0520 ± 0.4108	331285.7143 ± 130550.1702	70.0%	81/275
Q-learning ($\rho = 0.05, \epsilon = 0.75$)	4.27226 ± 0.0507	1399634.9000 ± 11249.3949	0.0%	0.0064 ± 0.0032	2182.1000 ± 1130.1038	nan ± nan	nan ± nan	0.0%	48/90
Q-learning ($\rho = 0.1, \epsilon = 0.75$)	3.70340 ± 0.0442	1214265.8000 ± 7179.8259	0.0%	0.0149 ± 0.0127	4909.4000 ± 4066.7467	2.6810 ± 0.3665	881222.2222 ± 119554.5229	90.0%	32/102
Q-learning ($\rho = 0.3, \epsilon = 0.75$)	3.82740 ± 0.0634	1242230.3000 ± 13564.5511	0.0%	0.0088 ± 0.0052	2930.9000 ± 1760.8187	2.1078 ± 0.8438	686857.1429 ± 277170.1018	70.0%	41/166
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	4.48964 ± 0.0547	1462640.9000 ± 15725.0141	0.0%	0.0069 ± 0.0031	2415.4000 ± 1131.9916	nan ± nan	nan ± nan	0.0%	70/317
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	5.41929 ± 0.0774	1769062.4000 ± 12003.8286	0.0%	0.0074 ± 0.0041	2355.2000 ± 1365.1223	nan ± nan	nan ± nan	0.0%	69/228
Q-learning ($\rho = 0.05, \epsilon = 0.9$)	7.48892 ± 0.1016	2508777.8000 ± 25977.4317	0.0%	0.0062 ± 0.0033	2158.8000 ± 1161.8588	6.4169 ± 0.5565	2151200.0000 ± 185532.0996	50.0%	48/84
Q-learning ($\rho = 0.1, \epsilon = 0.9$)	7.37564 ± 0.0839	2462495.5000 ± 30284.0138	0.0%	0.0085 ± 0.0082	2984.9000 ± 2743.6969	4.2728 ± 1.4063	1426800.0000 ± 473715.9064	100.0%	43/96
Q-learning ($\rho = 0.3, \epsilon = 0.9$)	8.10403 ± 0.1603	2709057.6000 ± 41297.9039	0.0%	0.0064 ± 0.0023	2127.9000 ± 887.1851	nan ± nan	nan ± nan	0.0%	41/127
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	9.13370 ± 0.1395	3059909.5000 ± 31495.7896	0.0%	0.0067 ± 0.0041	2185.0000 ± 1329.1192	nan ± nan	nan ± nan	0.0%	53/323
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	10.22223 ± 0.1653	3415851.1000 ± 36943.7384	0.0%	0.0093 ± 0.0057	3265.0000 ± 1934.7822	nan ± nan	nan ± nan	0.0%	85/381
Q-learning ($\rho = 0.05, \epsilon = 1$)	16.79910 ± 0.1905	5719151.1000 ± 58057.9365	0.0%	0.0043 ± 0.0021	1500.0000 ± 821.7840	13.2550 ± 2.0974	4527000.0000 ± 724314.5035	40.0%	46/80
Q-learning ($\rho = 0.1, \epsilon = 1$)	16.70501 ± 0.1752	5693211.2000 ± 51659.8073	0.0%	0.0060 ± 0.0078	1971.9000 ± 2369.6760	11.3464 ± 4.3977	3842000.0000 ± 1474000.0000	20.0%	40/81
Q-learning ($\rho = 0.3, \epsilon = 1$)	16.81030 ± 0.1995	5718518.9000 ± 50499.9419	0.0%	0.0058 ± 0.0033	1854.1000 ± 1014.0778	nan ± nan	nan ± nan	0.0%	51/102
Q-learning ($\rho = 0.5, \epsilon = 1$)	16.84281 ± 0.2543	5729605.4000 ± 67141.3314	0.0%	0.0127 ± 0.0117	4527.9000 ± 4167.9568	nan ± nan	nan ± nan	0.0%	61/275
Q-learning ($\rho = 0.7, \epsilon = 1$)	16.70733 ± 0.1561	5676013.0000 ± 51046.4526	0.0%	0.0059 ± 0.0042	1975.9000 ± 1363.3594	nan ± nan	nan ± nan	0.0%	40/211
Async Value Iteration	0.93243 ± 0.0116	1359328.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	42/75
Value Iteration	1.38258 ± 0.0123	2013312.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	48/66

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

Algorithm (Problem 11) ($\gamma = 0.7$)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.05, \epsilon = 0$)	1.85072 ± 0.0295	571158.4000 ± 1756.1736	0.0%	0.0075 ± 0.0034	2390.2000 ± 1008.7788	nan ± nan	nan ± nan	0.0%	29/285
Q-learning ($\rho = 0.1, \epsilon = 0$)	1.14824 ± 0.0237	356490.7000 ± 1679.6158	0.0%	0.0078 ± 0.0033	2359.9000 ± 993.0394	0.9849 ± 0.0242	305800.0000 ± 7639.3717	100.0%	31/41
Q-learning ($\rho = 0.3, \epsilon = 0$)	0.65108 ± 0.0161	202609.4000 ± 3396.0761	0.0%	0.0069 ± 0.0023	2246.9000 ± 733.8597	0.3335 ± 0.0299	104000.0000 ± 9121.4034	100.0%	29/42
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.58352 ± 0.0212	176486.2000 ± 5260.8166	0.0%	0.0078 ± 0.0041	2234.4000 ± 986.1301	0.2652 ± 0.0907	81600.0000 ± 25865.8075	100.0%	31/62
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.55968 ± 0.0267	172331.0000 ± 5244.1676	0.0%	0.0081 ± 0.0032	2443.8000 ± 1067.0162	0.2965 ± 0.1341	91000.0000 ± 42092.7547	100.0%	28/52
Q-learning ($\rho = 0.05, \epsilon = 0.25$)	1.87500 ± 0.0236	590310.7000 ± 2177.8359	0.0%	0.0095 ± 0.0053	3295.1000 ± 2025.5409	nan ± nan	nan ± nan	0.0%	28/65
Q-learning ($\rho = 0.1, \epsilon = 0.25$)	1.24466 ± 0.0400	389756.7000 ± 1779.5030	0.0%	0.0091 ± 0.0033	2694.3000 ± 919.2288	1.0451 ± 0.0761	327000.0000 ± 15588.4573	100.0%	25/57
Q-learning ($\rho = 0.3, \epsilon = 0.25$)	0.80251 ± 0.0194	252618.3000 ± 1626.5426	0.0%	0.0068 ± 0.0034	2190.8000 ± 1129.2834	0.3748 ± 0.0392	119300.0000 ± 12720.4560	100.0%	29/118
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.78585 ± 0.0135	249014.0000 ± 1344.5959	0.0%	0.0066 ± 0.0030	2190.0000 ± 799.4007	0.2369 ± 0.0311	75900.0000 ± 10520.9315	100.0%	29/75
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.88708 ± 0.0174	275623.1000 ± 1277.1430	0.0%	0.0069 ± 0.0042	2172.8000 ± 1219.5270	0.2573 ± 0.1515	80100.0000 ± 46601.3948	100.0%	37/192
Q-learning ($\rho = 0.05, \epsilon = 0.5$)	2.03455 ± 0.0377	655294.6000 ± 2854.5637	0.0%	0.0078 ± 0.0031	2561.8000 ± 1047.6895	nan ± nan	nan ± nan	0.0%	29/122
Q-learning ($\rho = 0.1, \epsilon = 0.5$)	1.46130 ± 0.0131	472283.0000 ± 953.4786	0.0%	0.0099 ± 0.0064	3090.8000 ± 1844.6877	1.1263 ± 0.0681	365800.0000 ± 24770.1433	100.0%	31/44
Q-learning ($\rho = 0.3, \epsilon = 0.5$)	1.09836 ± 0.0206	354386.5000 ± 1866.4678	0.0%	0.0087 ± 0.0035	2771.8000 ± 1181.1545	0.3983 ± 0.0440	130500.0000 ± 13522.2040	100.0%	30/46
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	1.13065 ± 0.0174	361634.5000 ± 1266.2115	0.0%	0.0076 ± 0.0041	2618.4000 ± 1386.6625	0.2841 ± 0.0807	92900.0000 ± 24320.5674	100.0%	36/142
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	1.25506 ± 0.0138	404392.1000 ± 2022.1663	0.0%	0.0064 ± 0.0029	1990.8000 ± 1007.2874	0.2065 ± 0.0271	67600.0000 ± 9624.9675	100.0%	38/368
Q-learning ($\rho = 0.05, \epsilon = 0.75$)	2.69296 ± 0.0191	892406.4000 ± 5942.4291	0.0%	0.0112 ± 0.0073	3719.5000 ± 2462.0602	2.4583 ± 0.1202	815444.4444 ± 41131.5265	90.0%	27/57
Q-learning ($\rho = 0.1, \epsilon = 0.75$)	2.25037 ± 0.0323	751469.6000 ± 3122.7089	0.0%	0.0077 ± 0.0029	2562.0000 ± 1019.4048	1.2703 ± 0.1051	425400.0000 ± 36540.9359	100.0%	28/44
Q-learning ($\rho = 0.3, \epsilon = 0.75$)	2.03944 ± 0.0189	679102.1000 ± 4756.2187	0.0%	0.0140 ± 0.0067	4856.2000 ± 2354.3572	0.4792 ± 0.0430	164200.0000 ± 15548.6334	100.0%	26/49
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	2.17193 ± 0.0318	715946.6000 ± 6384.1643	0.0%	0.0097 ± 0.0062	3138.9000 ± 1924.8360	0.4323 ± 0.1421	144800.0000 ± 47509.5780	100.0%	33/63
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	2.42936 ± 0.0384	801925.5000 ± 8217.3459	0.0%	0.0094 ± 0.0053	3023.2000 ± 1541.6154	0.4701 ± 0.1764	158300.0000 ± 58027.6658	100.0%	35/100
Q-learning ($\rho = 0.05, \epsilon = 0.9$)	4.93538 ± 0.0688	1663924.7000 ± 19238.1992	0.0%	0.0065 ± 0.0052	1999.6000 ± 1381.8728	2.9590 ± 0.1871	1000700.0000 ± 61799.7573	100.0%	31/55
Q-learning ($\rho = 0.1, \epsilon = 0.9$)	4.80960 ± 0.0639	1626720.0000 ± 21414.2769	0.0%	0.0060 ± 0.0050	2086.5000 ± 1628.0246	1.5324 ± 0.1175	520800.0000 ± 40703.3168	100.0%	27/44
Q-learning ($\rho = 0.3, \epsilon = 0.9$)	4.86495 ± 0.0493	1651366.2000 ± 15328.2049	0.0%	0.0057 ± 0.0036	1932.0000 ± 1066.7638	0.6609 ± 0.1607	228100.0000 ± 57118.2108	100.0%	27/58
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	5.22074 ± 0.1232	1750767.3000 ± 14806.5850	0.0%	0.0106 ± 0.0045	3021.4000 ± 1571.2568	0.8291 ± 0.3267	279900.0000 ± 110134.8719	100.0%	32/91
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	5.65502 ± 0.0808	1912059.2000 ± 18695.1183	0.0%	0.0048 ± 0.0029	1601.6000 ± 1027.8246	2.2342 ± 0.0000	759000.0000 ± 0.0000	10.0%	33/88
Q-learning ($\rho = 0.05, \epsilon = 1$)	15.41285 ± 0.0935	5318939.5000 ± 38522.8975	0.0%	0.0071 ± 0.0047	2445.2000 ± 1689.7547	6.1668 ± 1.4009	2127100.0000 ± 485526.4050	100.0%	27/51
Q-learning ($\rho = 0.1, \epsilon = 1$)	15.39910 ± 0.1935	5310918.5000 ± 52614.4491	0.0%	0.0065 ± 0.0060	2343.2000 ± 2314.2856	3.3551 ± 0.5765	1156000.0000 ± 191039.2630	100.0%	25/46
Q-learning ($\rho = 0.3, \epsilon = 1$)	15.47657 ± 0.1695	5346496.6000 ± 52074.0087	0.0%	0.0070 ± 0.0131	2170.3000 ± 4115.2852	1.5341 ± 0.4021	533700.0000 ± 141861.9399	100.0%	32/52
Q-learning ($\rho = 0.5, \epsilon = 1$)	15.47259 ± 0.2803	5334627.2000 ± 64805.3006	0.0%	0.0066 ± 0.0032	2404.2000 ± 1023.9160	5.0898 ± 3.5793	1761285.7143 ± 1225453.5853	70.0%	26/97
Q-learning ($\rho = 0.7, \epsilon = 1$)	15.40955 ± 0.1712	5329222.6000 ± 49331.8510	0.0%	0.0094 ± 0.0073	3404.0000 ± 2618.4177	nan ± nan	nan ± nan	0.0%	32/133
Async Value Iteration	0.39380 ± 0.0078	582080.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	31/48
Value Iteration	0.61098 ± 0.0218	866272.0000 ± 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 ρ in stochastic problem with $\gamma = 0.9$ (Problem 11).

Algorithm (Problem 11) ($\gamma = 0.9$)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.05, \epsilon = 0$)	1.27863 ± 0.0220	402517.4000 ± 1531.6856	0.0%	0.0063 ± 0.0020	1962.7000 ± 609.9456	1.2686 ± 0.0234	399500.0000 ± 3278.7193	40.0%	25/622
Q-learning ($\rho = 0.1, \epsilon = 0$)	0.78346 ± 0.0149	244012.8000 ± 1006.6520	0.0%	0.0056 ± 0.0013	1728.4000 ± 349.4402	0.6537 ± 0.0220	204400.0000 ± 5986.6518	100.0%	23/32
Q-learning ($\rho = 0.3, \epsilon = 0$)	0.42620 ± 0.0100	134923.5000 ± 1267.5736	0.0%	0.0053 ± 0.0012	1692.2000 ± 430.9220	0.2215 ± 0.0144	70400.0000 ± 3411.7444	100.0%	22/38
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.37307 ± 0.0201	115730.1000 ± 3152.5928	0.0%	0.0059 ± 0.0017	1798.7000 ± 519.0761	0.1796 ± 0.0522	54000.0000 ± 16983.6523	90.0%	26/32
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.35158 ± 0.0103	109529.8000 ± 2347.4768	0.0%	0.0076 ± 0.0061	1675.4000 ± 436.1360	0.1594 ± 0.0886	49800.0000 ± 27128.5827	100.0%	22/31
Q-learning ($\rho = 0.05, \epsilon = 0.25$)	1.31733 ± 0.0158	421869.2000 ± 1067.8852	0.0%	0.0072 ± 0.0021	2341.6000 ± 601.9774	1.3034 ± 0.0097	418000.0000 ± 4301.1626	40.0%	23/29
Q-learning ($\rho = 0.1, \epsilon = 0.25$)	0.84795 ± 0.0137	273211.9000 ± 747.9059	0.0%	0.0060 ± 0.0037	1993.6000 ± 949.0605	0.6631 ± 0.0242	213800.0000 ± 8518.2158	100.0%	22/30
Q-learning ($\rho = 0.3, \epsilon = 0.25$)	0.52191 ± 0.0093	168343.1000 ± 603.8706	0.0%	0.0065 ± 0.0023	1990.0000 ± 605.5198	0.2346 ± 0.0223	76400.0000 ± 6843.9755	100.0%	22/33
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.47103 ± 0.0143	150304.9000 ± 799.3068	0.0%	0.0048 ± 0.0018	1637.8000 ± 494.4348	0.1729 ± 0.0353	55300.0000 ± 10344.5638	100.0%	22/30
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.48917 ± 0.0140	153594.8000 ± 1154.1807	0.0%	0.0065 ± 0.0008	1936.3000 ± 387.2782	0.1570 ± 0.0340	49100.0000 ± 10074.2245	100.0%	23/161
Q-learning ($\rho = 0.05, \epsilon = 0.5$)	1.43533 ± 0.0133	472931.3000 ± 1386.4713	0.0%	0.0080 ± 0.0028	2720.3000 ± 898.2796	1.3476 ± 0.0289	443600.0000 ± 8879.1892	100.0%	22/29
Q-learning ($\rho = 0.1, \epsilon = 0.5$)	1.02841 ± 0.0125	337065.6000 ± 816.3183	0.0%	0.0088 ± 0.0023	3119.2000 ± 882.9344	0.6871 ± 0.0226	226000.0000 ± 5692.0998	100.0%	23/33
Q-learning ($\rho = 0.3, \epsilon = 0.5$)	0.74511 ± 0.0119	242345.7000 ± 950.0345	0.0%	0.0082 ± 0.0024	2677.6000 ± 785.5479	0.2627 ± 0.0294	85800.0000 ± 7547.1849	100.0%	23/30
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.68277 ± 0.0106	224632.0000 ± 818.2979	0.0%	0.0076 ± 0.0025	2580.5000 ± 770.9068	0.1686 ± 0.0207	56000.0000 ± 7183.3140	100.0%	22/37
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.69505 ± 0.0094	227732.5000 ± 1397.4116	0.0%	0.0064 ± 0.0017	2167.3000 ± 607.9712	0.1483 ± 0.0341	49300.0000 ± 10982.2584	100.0%	24/30
Q-learning ($\rho = 0.05, \epsilon = 0.75$)	1.93422 ± 0.0228	654523.0000 ± 2364.1268	0.0%	0.0099 ± 0.0044	3357.4000 ± 1379.9961	1.5492 ± 0.0578	524500.0000 ± 13843.7712	100.0%	22/36
Q-learning ($\rho = 0.1, \epsilon = 0.75$)	1.63222 ± 0.0211	549482.2000 ± 2786.6671	0.0%	0.0081 ± 0.0021	2814.4000 ± 908.8171	0.7917 ± 0.0435	268300.0000 ± 13520.7248	100.0%	22/29
Q-learning ($\rho = 0.3, \epsilon = 0.75$)	1.42428 ± 0.0276	476723.7000 ± 3385.8993	0.0%	0.0080 ± 0.0045	2788.3000 ± 1691.7654	0.2968 ± 0.0298	102700.0000 ± 9839.2073	100.0%	22/32
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	1.38091 ± 0.0135	465512.0000 ± 3544.0541	0.0%	0.0079 ± 0.0020	2541.1000 ± 597.5407	0.2098 ± 0.0215	70100.0000 ± 6122.9078	100.0%	23/29
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	1.41781 ± 0.0196	475307.6000 ± 2213.8200	0.0%	0.0082 ± 0.0037	2767.1000 ± 949.0730	0.1819 ± 0.0238	59700.0000 ± 7823.6820	100.0%	22/66
Q-learning ($\rho = 0.05, \epsilon = 0.9$)	3.70833 ± 0.0695	1272329.3000 ± 8113.6820	0.0%	0.0083 ± 0.0059	2779.2000 ± 1605.7997	1.9743 ± 0.1360	681400.0000 ± 51734.3213	100.0%	22/28
Q-learning ($\rho = 0.1, \epsilon = 0.9$)	3.59731 ± 0.0593	1233127.9000 ± 14698.8456	0.0%	0.0071 ± 0.0050	2438.2000 ± 1678.0681	0.9816 ± 0.0597	339900.0000 ± 19206.5093	100.0%	22/40
Q-learning ($\rho = 0.3, \epsilon = 0.9$)	3.50814 ± 0.0535	1210221.3000 ± 10184.3410	0.0%	0.0059 ± 0.0023	2050.8000 ± 907.8454	0.3729 ± 0.0360	130100.0000 ± 12397.1771	100.0%	22/29
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	3.53971 ± 0.0240	1218752.1000 ± 11656.3795	0.0%	0.0083 ± 0.0054	2860.3000 ± 2156.4909	0.2771 ± 0.0399	95800.0000 ± 14743.1340	100.0%	23/30
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	3.59111 ± 0.0557	1236672.1000 ± 9247.5373	0.0%	0.0065 ± 0.0052	2305.2000 ± 1751.9880	0.2315 ± 0.0515	80700.0000 ± 17675.1238	100.0%	23/48
Q-learning ($\rho = 0.05, \epsilon = 1$)	14.25785 ± 0.1472	4990352.9000 ± 31912.7481	0.0%	0.0049 ± 0.0041	1902.8000 ± 1634.7228	3.5919 ± 0.3768	1260400.0000 ± 122172.1736	100.0%	22/35
Q-learning ($\rho = 0.1, \epsilon = 1$)	14.13757 ± 0.1038	4960982.4000 ± 33989.7238	0.0%	0.0055 ± 0.0048	1922.8000 ± 1818.9051	1.8594 ± 0.3989	656800.0000 ± 138768.7285	100.0%	23/29
Q-learning ($\rho = 0.3, \epsilon = 1$)	14.24336 ± 0.1430	5003448.1000 ± 36205.8142	0.0%	0.0053 ± 0.0034	1959.2000 ± 1138.7284	0.6274 ± 0.1134	224100.0000 ± 40515.3057	100.0%	22/28
Q-learning ($\rho = 0.5, \epsilon = 1$)	14.17406 ± 0.1845	4979801.0000 ± 44084.7961	0.0%	0.0079 ± 0.0083	2509.2000 ± 2684.7604	0.4008 ± 0.0904	142000.0000 ± 30106.4777	100.0%	24/28
Q-learning ($\rho = 0.7, \epsilon = 1$)	14.28713 ± 0.2073	4996836.4000 ± 53793.4112	0.0%	0.0066 ± 0.0072	2100.7000 ± 1831.7205	0.3475 ± 0.0412	123600.0000 ± 14242.1908	100.0%	22/166
Async Value Iteration	0.19530 ± 0.0071	291040.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	23/31
Value Iteration	0.27897 ± 0.0099	410880.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	23/29

Table 22. Performance of DP methods and Q-learning as we change the learning rate ρ across Problems 1, 10, 11 with $\gamma = 0.5$ and ϵ-decay.

Problem	Algorithm ($\gamma = 0.5$)	Runtime (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
3*Problem 1	Q-learning ($\rho = 0.1$, ϵ-decay)	3.24582 ± 0.0481	975975.5000 ± 12459.5353	0.0%	0.0074 ± 0.0049	2509.1000 ± 1566.0333	3.2671 ± 0.0000	944000.0000 ± 0.0000	10.0%	50/128
	Q-learning ($\rho = 0.5$, ϵ-decay)	4.10926 ± 0.0812	1232561.3000 ± 24785.4891	0.0%	0.0060 ± 0.0070	1741.4000 ± 1814.7200	3.7953 ± 0.1545	1144700.0000 ± 44643.1406	100.0%	67/229
	Q-learning ($\rho = 0.7$, ϵ-decay)	4.79570 ± 0.0414	1436071.3000 ± 10869.9424	0.0%	0.0063 ± 0.0049	2151.4000 ± 1639.6912	4.4272 ± 0.0812	1331111.1111 ± 25653.4357	90.0%	78/221
3*Problem 10	Q-learning ($\rho = 0.1$, ϵ-decay)	8.58504 ± 0.0795	2559279.9000 ± 8432.7009	0.0%	0.1930 ± 0.1375	60781.8000 ± 43576.0453	8.2682 ± 0.1506	2470200.0000 ± 50469.3967	100.0%	126/239
	Q-learning ($\rho = 0.5$, ϵ-decay)	10.73371 ± 0.1061	3188840.5000 ± 17703.9743	0.0%	0.1896 ± 0.1667	59183.9000 ± 52306.6374	9.2165 ± 0.1096	2762800.0000 ± 31054.1463	100.0%	210/399
	Q-learning ($\rho = 0.7$, ϵ-decay)	12.64540 ± 0.1723	3698278.3000 ± 22997.6596	0.0%	0.2308 ± 0.1973	70310.9000 ± 59964.0379	11.1641 ± 0.2273	3285444.4444 ± 37706.7961	90.0%	231/686
3*Problem 11	Q-learning ($\rho = 0.1$, ϵ-decay)	3.51188 ± 0.0792	1034769.7000 ± 17122.5113	0.0%	0.0123 ± 0.0097	3871.0000 ± 3051.3563	2.9562 ± 0.1796	879600.0000 ± 47417.7182	100.0%	44/82
	Q-learning ($\rho = 0.5$, ϵ-decay)	4.43853 ± 0.0516	1300652.8000 ± 9965.3303	0.0%	0.0070 ± 0.0055	1889.8000 ± 1337.9283	3.6388 ± 0.1191	1080200.0000 ± 33969.9868	100.0%	47/162
	Q-learning ($\rho = 0.7$, ϵ-decay)	5.15350 ± 0.0625	1510802.2000 ± 22221.0360	0.0%	0.0101 ± 0.0100	3000.6000 ± 2657.1842	4.4489 ± 0.1049	1316400.0000 ± 29127.9934	100.0%	69/260

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

Problem	Algorithm ($\gamma = 0.7$)	Runtime (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
3*Problem 1	Q-learning ($\rho = 0.1$, ϵ-decay)	2.16270 ± 0.0419	658836.4000 ± 9598.6608	0.0%	0.0067 ± 0.0069	2038.3000 ± 2274.6604	2.1245 ± 0.0848	644400.0000 ± 16740.3704	50.0%	36/52
	Q-learning ($\rho = 0.5$, ϵ-decay)	2.33761 ± 0.0476	714369.3000 ± 10360.6553	0.0%	0.0057 ± 0.0046	1664.8000 ± 1389.9136	2.2245 ± 0.0736	682400.0000 ± 15007.9979	100.0%	43/89
	Q-learning ($\rho = 0.7$, ϵ-decay)	2.54275 ± 0.0339	779184.0000 ± 9794.5332	0.0%	0.0087 ± 0.0050	2668.1000 ± 1413.2517	2.3245 ± 0.0358	716100.0000 ± 12856.5159	100.0%	40/964
3*Problem 10	Q-learning ($\rho = 0.1$, ϵ-decay)	5.95362 ± 0.2101	1740023.6000 ± 12232.4373	0.0%	0.0666 ± 0.0791	20799.5000 ± 25673.8536	5.6562 ± 0.3261	1657250.0000 ± 72912.1903	40.0%	90/117
	Q-learning ($\rho = 0.5$, ϵ-decay)	6.25790 ± 0.1271	1890867.9000 ± 19252.7588	0.0%	0.1419 ± 0.1046	44964.1000 ± 33965.3956	5.2121 ± 0.2097	1590800.0000 ± 48186.7202	100.0%	106/221
	Q-learning ($\rho = 0.7$, ϵ-decay)	6.95639 ± 0.1001	2080745.9000 ± 11179.6210	0.0%	0.1271 ± 0.0869	38293.5000 ± 25260.4927	5.6864 ± 0.2002	1720100.0000 ± 52605.0378	100.0%	143/252
3*Problem 11	Q-learning ($\rho = 0.1$, ϵ-decay)	2.30780 ± 0.0611	696267.3000 ± 15229.0235	0.0%	0.0065 ± 0.0044	1979.8000 ± 1376.0974	1.7137 ± 0.0921	525600.0000 ± 27372.2487	100.0%	24/52
	Q-learning ($\rho = 0.5$, ϵ-decay)	2.50699 ± 0.0438	746330.8000 ± 15952.4150	0.0%	0.0069 ± 0.0062	2160.8000 ± 1684.5450	1.2700 ± 0.4061	387800.0000 ± 121489.7527	100.0%	34/60
	Q-learning ($\rho = 0.7$, ϵ-decay)	2.87144 ± 0.0884	830611.5000 ± 12281.4056	0.0%	0.0101 ± 0.0097	2863.2000 ± 3077.0052	1.9816 ± 0.3205	583300.0000 ± 88221.3693	100.0%	35/195

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

Problem	Algorithm ($\gamma = 0.9$)	Runtime (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
3*Problem 1	Q-learning ($\rho = 0.1$, ϵ-decay)	1.66952 ± 0.0451	508233.7000 ± 6252.7432	0.0%	0.0067 ± 0.0072	1956.8000 ± 2326.8822	1.5912 ± 0.0565	488800.0000 ± 14161.9208	50.0%	28/42
	Q-learning ($\rho = 0.5$, ϵ-decay)	1.66440 ± 0.0528	515043.7000 ± 11570.6701	0.0%	0.0061 ± 0.0057	2005.3000 ± 1886.0563	1.6242 ± 0.0713	506000.0000 ± 19869.5747	50.0%	29/37
	Q-learning ($\rho = 0.7$, ϵ-decay)	1.69829 ± 0.0224	524086.2000 ± 7062.3417	0.0%	0.0076 ± 0.0045	2102.8000 ± 1467.6715	1.5281 ± 0.1464	475900.0000 ± 43000.0000	100.0%	29/169
3*Problem 10	Q-learning ($\rho = 0.1$, ϵ-decay)	4.45455 ± 0.1874	1352699.6000 ± 12188.7971	0.0%	0.0929 ± 0.0452	29464.5000 ± 14158.1076	4.0246 ± 0.3538	1227500.0000 ± 69655.2223	100.0%	71/83
	Q-learning ($\rho = 0.5$, ϵ-decay)	4.57553 ± 0.1357	1359802.1000 ± 8836.6308	0.0%	0.1460 ± 0.0860	45129.2000 ± 26601.8713	2.8682 ± 0.9407	868800.0000 ± 278771.5193	100.0%	73/267
	Q-learning ($\rho = 0.7$, ϵ-decay)	4.80031 ± 0.2351	1404558.4000 ± 16347.0547	0.0%	0.1590 ± 0.0824	48680.1000 ± 25518.9627	3.6637 ± 0.4287	1080400.0000 ± 98097.0948	100.0%	91/176
3*Problem 11	Q-learning ($\rho = 0.1$, ϵ-decay)	1.71151 ± 0.0329	527798.8000 ± 5334.1851	0.0%	0.0109 ± 0.0088	3629.6000 ± 2917.8036	1.1859 ± 0.0716	369800.0000 ± 20183.1613	100.0%	22/28
	Q-learning ($\rho = 0.5$, ϵ-decay)	1.72270 ± 0.0276	529684.5000 ± 7678.3650	0.0%	0.0054 ± 0.0055	1691.9000 ± 1648.9380	0.4592 ± 0.0760	144000.0000 ± 22834.1849	100.0%	22/36
	Q-learning ($\rho = 0.7$, ϵ-decay)	1.80766 ± 0.0332	549734.0000 ± 9597.4643	0.0%	0.0054 ± 0.0031	1609.6000 ± 898.9969	0.3502 ± 0.0692	111800.0000 ± 22964.3202	100.0%	22/261

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

Algorithm (Problem 8) ($\gamma = 0.5$)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.1, \epsilon = 0$)	0.13995 ± 0.0056	44668.1000 ± 954.4866	0.0%	0.0010 ± 0.0000	60.2000 ± 24.3343	0.0371 ± 0.0107	12000.0000 ± 3521.3634	100.0%	7/36
Q-learning ($\rho = 0.1, \epsilon = 0.25$)	0.17486 ± 0.0069	56086.3000 ± 461.1481	0.0%	0.0010 ± 0.0000	71.0000 ± 46.9766	0.0349 ± 0.0166	11800.0000 ± 5741.0800	100.0%	6/21
Q-learning ($\rho = 0.1, \epsilon = 0.5$)	0.22090 ± 0.0060	74276.3000 ± 451.3032	0.0%	nan ± nan	60.9000 ± 45.1762	0.0318 ± 0.0132	10800.0000 ± 4686.1498	100.0%	8/21
Q-learning ($\rho = 0.1, \epsilon = 0.75$)	0.31864 ± 0.0099	107960.5000 ± 1346.5570	0.0%	0.0011 ± 0.0001	58.3000 ± 57.5379	0.0471 ± 0.0234	16100.0000 ± 7542.5460	100.0%	8/40
Q-learning ($\rho = 0.1, \epsilon = 0.9$)	0.41847 ± 0.0080	144128.9000 ± 1953.6720	0.0%	0.0010 ± 0.0000	69.9000 ± 37.6708	0.0474 ± 0.0300	16300.0000 ± 9808.6696	100.0%	7/67
Q-learning ($\rho = 0.1, \epsilon = 1$)	0.53000 ± 0.0215	181779.4000 ± 1735.9936	0.0%	0.0010 ± 0.0000	72.5000 ± 30.3883	0.0548 ± 0.0155	19000.0000 ± 5549.7748	100.0%	6/31
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0$)	0.14912 ± 0.0099	43173.7000 ± 579.2908	0.0%	0.0010 ± 0.0000	70.7000 ± 31.8467	0.0861 ± 0.0151	24857.1429 ± 5890.1509	70.0%	9/27
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0.25$)	0.17867 ± 0.0057	52892.4000 ± 679.6827	0.0%	0.0010 ± 0.0000	71.1000 ± 53.0367	0.1180 ± 0.0425	34400.0000 ± 12354.7562	50.0%	8/28
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0.5$)	0.22891 ± 0.0133	69617.6000 ± 3001.1876	10.0%	0.0010 ± 0.0000	89.0000 ± 42.8416	0.1179 ± 0.0322	36666.6667 ± 10811.5165	60.0%	6/25
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0.75$)	0.33659 ± 0.0047	104463.0000 ± 1254.3633	0.0%	0.0013 ± 0.0005	94.2000 ± 50.7756	0.1568 ± 0.0468	48333.3333 ± 14290.6341	60.0%	12/24
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0.9$)	0.42532 ± 0.0468	133795.0000 ± 16046.5053	20.0%	0.0010 ± 0.0000	80.4000 ± 59.3013	0.2706 ± 0.0882	84875.0000 ± 27342.4464	80.0%	7/28
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 1$)	0.56828 ± 0.0132	181020.5000 ± 2239.2016	0.0%	0.0010 ± 0.0000	44.0000 ± 22.7420	0.3087 ± 0.0925	98500.0000 ± 28860.2957	60.0%	9/24

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

Algorithm (Problem 8) ($\gamma = 0.7$)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.1, \epsilon = 0$)	0.08366 ± 0.0022	26825.4000 ± 129.0102	0.0%	0.0010 ± 0.0000	55.1000 ± 10.0941	0.0163 ± 0.0089	5200.0000 ± 27.49.5454	100.0%	6/11
Q-learning ($\rho = 0.1, \epsilon = 0.25$)	0.10487 ± 0.0029	34645.2000 ± 337.1219	0.0%	nan ± nan	54.4000 ± 17.3966	0.0167 ± 0.0029	5400.0000 ± 1113.5529	100.0%	7/17
Q-learning ($\rho = 0.1, \epsilon = 0.5$)	0.14780 ± 0.0060	48376.7000 ± 404.5768	0.0%	0.0010 ± 0.0000	77.5000 ± 36.2939	0.0250 ± 0.0260	8400.0000 ± 8357.0330	100.0%	6/12
Q-learning ($\rho = 0.1, \epsilon = 0.75$)	0.22654 ± 0.0099	76800.8000 ± 716.2207	0.0%	0.0014 ± 0.0005	82.0000 ± 45.4026	0.0261 ± 0.0201	8900.0000 ± 6315.8531	100.0%	6/12
Q-learning ($\rho = 0.1, \epsilon = 0.9$)	0.32994 ± 0.0105	113612.1000 ± 1430.1477	0.0%	0.0010 ± 0.0000	29.8000 ± 14.6820	0.0237 ± 0.0070	8000.0000 ± 2607.6810	100.0%	6/11
Q-learning ($\rho = 0.1, \epsilon = 1$)	0.46176 ± 0.0121	162542.8000 ± 1633.1799	0.0%	0.0010 ± 0.0000	43.9000 ± 32.7275	0.0309 ± 0.0200	10900.0000 ± 6655.0733	100.0%	6/13
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0$)	0.09052 ± 0.0108	24763.2000 ± 2091.1724	20.0%	0.0010 ± 0.0000	54.4000 ± 15.1142	0.0280 ± 0.0176	7800.0000 ± 4770.7442	100.0%	6/15
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0.25$)	0.09668 ± 0.0242	28498.0000 ± 6484.9181	60.0%	0.0010 ± 0.0000	62.4000 ± 15.3571	0.0242 ± 0.0090	7000.0000 ± 2607.6810	100.0%	7/17
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0.5$)	0.12390 ± 0.0372	36596.0000 ± 9908.6227	60.0%	0.0010 ± 0.0000	64.5000 ± 23.4659	0.0254 ± 0.0073	7400.0000 ± 2244.9944	100.0%	6/15
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0.75$)	0.12229 ± 0.0471	38215.6000 ± 14606.2351	100.0%	0.0012 ± 0.0000	53.3000 ± 18.9581	0.0362 ± 0.0181	11400.0000 ± 5642.6944	100.0%	6/14
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0.9$)	0.22085 ± 0.0954	69579.4000 ± 29671.2468	90.0%	0.0010 ± 0.0000	62.9000 ± 36.2642	0.0435 ± 0.0146	13800.0000 ± 4955.8047	100.0%	6/11
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 1$)	0.31037 ± 0.1316	99554.0000 ± 41351.2545	80.0%	0.0009 ± 0.0000	41.7000 ± 27.8138	0.0644 ± 0.0255	20900.0000 ± 9027.1812	100.0%	6/12

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

Algorithm (Problem 8) ($\gamma = 0.9$)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.1, \epsilon = 0$)	0.06127 ± 0.0056	18524.3000 ± 54.7669	0.0%	0.0010 ± 0.0000	48.2000 ± 5.9967	0.0135 ± 0.0053	4100.0000 ± 1757.8396	100.0%	6/8
Q-learning ($\rho = 0.1, \epsilon = 0.25$)	0.08170 ± 0.0057	24880.8000 ± 140.5125	0.0%	0.0010 ± 0.0001	54.8000 ± 11.6944	0.0119 ± 0.0039	3700.0000 ± 1268.8578	100.0%	6/8
Q-learning ($\rho = 0.1, \epsilon = 0.5$)	0.11746 ± 0.0055	35674.7000 ± 455.7789	10.0%	0.0010 ± 0.0000	55.7000 ± 13.4242	0.0117 ± 0.0072	3500.0000 ± 1802.7756	100.0%	6/9
Q-learning ($\rho = 0.1, \epsilon = 0.75$)	0.15690 ± 0.0437	53220.4000 ± 14082.3169	20.0%	0.0010 ± 0.0000	59.2000 ± 24.9151	0.0108 ± 0.0033	3600.0000 ± 1200.0000	100.0%	6/8
Q-learning ($\rho = 0.1, \epsilon = 0.9$)	0.23116 ± 0.0650	79250.4000 ± 21104.3816	50.0%	0.0010 ± 0.0000	60.8000 ± 44.8415	0.0149 ± 0.0042	5000.0000 ± 1483.2397	100.0%	6/8
Q-learning ($\rho = 0.1, \epsilon = 1$)	0.35310 ± 0.1183	118078.8000 ± 40758.0824	40.0%	nan ± nan	37.7000 ± 28.8134	0.0140 ± 0.0057	4500.0000 ± 1746.4249	100.0%	6/9
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0$)	0.06102 ± 0.0045	17667.7000 ± 229.2322	10.0%	0.0012 ± 0.0001	46.6000 ± 4.5431	0.0058 ± 0.0026	1800.0000 ± 979.7959	100.0%	6/9
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0.25$)	0.03769 ± 0.0187	11380.3000 ± 5483.2880	90.0%	0.0010 ± 0.0000	53.6000 ± 8.9800	0.0072 ± 0.0032	2000.0000 ± 894.4272	100.0%	6/9
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0.5$)	0.03590 ± 0.0181	10505.7000 ± 5005.2690	100.0%	0.0010 ± 0.0000	60.4000 ± 14.4305	0.0128 ± 0.0056	3800.0000 ± 1400.0000	100.0%	6/8
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0.75$)	0.04336 ± 0.0375	14210.7000 ± 13020.7933	100.0%	0.0010 ± 0.0000	39.8000 ± 19.9690	0.0110 ± 0.0020	3400.0000 ± 663.3250	100.0%	6/9
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0.9$)	0.04518 ± 0.0210	14327.6000 ± 6430.9359	100.0%	0.0010 ± 0.0000	66.2000 ± 47.9183	0.0147 ± 0.0069	4600.0000 ± 2200.0000	100.0%	6/8
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 1$)	0.06014 ± 0.0207	18629.8000 ± 6211.6859	100.0%	0.0009 ± 0.0001	47.3000 ± 39.1077	0.0167 ± 0.0099	5300.0000 ± 3067.5723	100.0%	6/8

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

Algorithm (Problem 16) ($\gamma = 0.5$)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.1, \epsilon = 0$)	0.12082 ± 0.0100	41389.8000 ± 4158.8417	10.0%	0.0010 ± 0.0000	39.2000 ± 20.4929	0.0482 ± 0.0255	16375.0000 ± 8484.3606	80.0%	6/15
Q-learning ($\rho = 0.1, \epsilon = 0.25$)	0.15886 ± 0.0042	53968.8000 ± 620.2851	0.0%	0.0010 ± 0.0000	59.3000 ± 43.9342	0.0585 ± 0.0469	19600.0000 ± 15344.0542	100.0%	5/33
Q-learning ($\rho = 0.1, \epsilon = 0.5$)	0.18802 ± 0.0049	65456.1000 ± 720.4785	0.0%	nan ± nan	31.9000 ± 19.5164	0.0498 ± 0.0374	17300.0000 ± 13183.7021	100.0%	5/24
Q-learning ($\rho = 0.1, \epsilon = 0.75$)	0.23202 ± 0.0066	81139.0000 ± 1646.8644	0.0%	0.0010 ± 0.0000	28.6000 ± 12.8390	0.0478 ± 0.0484	16900.0000 ± 17397.9884	100.0%	10/56
Q-learning ($\rho = 0.1, \epsilon = 0.9$)	0.26826 ± 0.0107	94253.2000 ± 1388.8168	0.0%	nan ± nan	50.2000 ± 55.3711	0.0349 ± 0.0296	12100.0000 ± 10348.4298	100.0%	5/29
Q-learning ($\rho = 0.1, \epsilon = 1$)	0.29285 ± 0.0062	104646.4000 ± 2076.6652	0.0%	0.0011 ± 0.0001	30.6000 ± 15.9762	0.0648 ± 0.0501	23100.0000 ± 17026.1563	100.0%	7/28
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0$)	0.12767 ± 0.0217	38422.6000 ± 6241.5515	30.0%	0.0010 ± 0.0000	48.8000 ± 37.8042	0.0885 ± 0.0245	27000.0000 ± 7483.3148	60.0%	5/36
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0.25$)	0.13465 ± 0.0287	42266.5000 ± 9628.2294	50.0%	nan ± nan	26.5000 ± 8.8798	0.0957 ± 0.0302	29714.2857 ± 9779.1949	70.0%	5/28
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0.5$)	0.19607 ± 0.0106	61402.8000 ± 3189.0828	10.0%	nan ± nan	28.1000 ± 14.8624	0.1384 ± 0.0486	43750.0000 ± 15554.3402	80.0%	5/51
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0.75$)	0.21758 ± 0.0316	69623.3000 ± 10517.6685	50.0%	0.0010 ± 0.0000	44.2000 ± 31.1634	0.1212 ± 0.0389	39000.0000 ± 12247.4487	70.0%	5/17
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0.9$)	0.25670 ± 0.0721	80731.8000 ± 22291.8848	30.0%	nan ± nan	35.3000 ± 35.6147	0.1261 ± 0.0868	38571.4286 ± 23993.1963	70.0%	7/19
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 1$)	0.27503 ± 0.0483	90129.6000 ± 14557.5210	70.0%	0.0010 ± 0.0000	29.5000 ± 21.7083	0.1364 ± 0.0440	44500.0000 ± 15001.6666	100.0%	5/22

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

Algorithm (Problem 16) ($\gamma = 0.7$)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.1, \epsilon = 0$)	0.06361 ± 0.0184	19818.4000 ± 5344.0573	20.0%	nan ± nan	26.6000 ± 10.6414	0.0147 ± 0.0182	4500.0000 ± 5835.2378	100.0%	5/13
Q-learning ($\rho = 0.1, \epsilon = 0.25$)	0.08052 ± 0.0034	27563.9000 ± 281.9179	0.0%	0.0010 ± 0.0000	32.8000 ± 24.5226	0.0162 ± 0.0105	5600.0000 ± 3411.7444	100.0%	5/18
Q-learning ($\rho = 0.1, \epsilon = 0.5$)	0.10427 ± 0.0084	36515.4000 ± 2258.4497	10.0%	0.0010 ± 0.0000	26.3000 ± 13.4911	0.0242 ± 0.0080	8200.0000 ± 2712.9320	100.0%	5/21
Q-learning ($\rho = 0.1, \epsilon = 0.75$)	0.15310 ± 0.0062	53967.9000 ± 1433.2967	10.0%	0.0010 ± 0.0000	22.9000 ± 19.0916	0.0202 ± 0.0185	6800.0000 ± 6209.6699	100.0%	5/13
Q-learning ($\rho = 0.1, \epsilon = 0.9$)	0.20158 ± 0.0093	71726.7000 ± 857.7608	0.0%	0.0010 ± 0.0000	25.6000 ± 11.9348	0.0208 ± 0.0134	7400.0000 ± 4800.0000	100.0%	5/16
Q-learning ($\rho = 0.1, \epsilon = 1$)	0.24490 ± 0.0097	87916.0000 ± 1119.1048	0.0%	0.0010 ± 0.0000	21.8000 ± 11.2854	0.0143 ± 0.0083	5000.0000 ± 3033.1502	100.0%	5/10
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0$)	0.02469 ± 0.0162	7103.7000 ± 4949.0689	100.0%	0.0010 ± 0.0000	25.9000 ± 18.4578	0.0099 ± 0.0063	2800.0000 ± 1536.2291	100.0%	5/13
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0.25$)	0.02524 ± 0.0117	7805.6000 ± 3735.9932	100.0%	0.0010 ± 0.0000	23.6000 ± 8.3570	0.0068 ± 0.0031	2000.0000 ± 894.4272	100.0%	5/20
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0.5$)	0.02912 ± 0.0077	9208.7000 ± 2478.5133	100.0%	0.0008 ± 0.0002	23.9000 ± 8.0554	0.0153 ± 0.0100	4800.0000 ± 3249.6154	100.0%	5/17
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0.75$)	0.04472 ± 0.0277	14314.8000 ± 8905.4751	100.0%	0.0010 ± 0.0000	34.6000 ± 29.5845	0.0170 ± 0.0066	5400.0000 ± 2059.1260	100.0%	5/11
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0.9$)	0.04224 ± 0.0271	13520.0000 ± 9318.7356	100.0%	0.0010 ± 0.0000	24.4000 ± 19.0536	0.0223 ± 0.0210	7200.0000 ± 7263.6079	100.0%	6/14
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 1$)	0.04713 ± 0.0323	15417.2000 ± 10797.2421	100.0%	0.0010 ± 0.0000	31.3000 ± 28.4817	0.0126 ± 0.0060	4000.0000 ± 2190.8902	100.0%	6/16

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

Algorithm (Problem 16) ($\gamma = 0.9$)	Run time (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.1, \epsilon = 0$)	0.02624 ± 0.0189	8753.4000 ± 6058.6937	50.0%	nan ± nan	16.9000 ± 4.9890	0.0035 ± 0.0013	1300.0000 ± 458.2576	100.0%	5/7
Q-learning ($\rho = 0.1, \epsilon = 0.25$)	0.03005 ± 0.0181	10329.0000 ± 6478.4042	70.0%	nan ± nan	22.8000 ± 11.3384	0.0060 ± 0.0043	2100.0000 ± 1577.9734	100.0%	5/9
Q-learning ($\rho = 0.1, \epsilon = 0.5$)	0.03391 ± 0.0151	11905.3000 ± 5261.5667	100.0%	0.0010 ± 0.0000	17.1000 ± 4.4822	0.0073 ± 0.0057	2600.0000 ± 2107.1308	100.0%	5/7
Q-learning ($\rho = 0.1, \epsilon = 0.75$)	0.04512 ± 0.0313	15109.6000 ± 10382.8197	100.0%	nan ± nan	17.7000 ± 8.9560	0.0084 ± 0.0057	2800.0000 ± 1886.7962	100.0%	5/7
Q-learning ($\rho = 0.1, \epsilon = 0.9$)	0.04857 ± 0.0329	18009.6000 ± 12123.9792	100.0%	0.0010 ± 0.0000	20.3000 ± 10.2864	0.0058 ± 0.0028	2100.0000 ± 1044.0307	100.0%	5/7
Q-learning ($\rho = 0.1, \epsilon = 1$)	0.04663 ± 0.0486	16819.9000 ± 17603.7841	100.0%	0.0007 ± 0.0003	25.0000 ± 18.4120	0.0111 ± 0.0077	3700.0000 ± 2282.5424	100.0%	5/9
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0$)	0.00990 ± 0.0054	3102.0000 ± 1758.0113	100.0%	0.0011 ± 0.0000	17.3000 ± 8.0131	0.0056 ± 0.0025	1700.0000 ± 781.0250	100.0%	5/10
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0.25$)	0.00769 ± 0.0041	2203.3000 ± 1165.3702	100.0%	nan ± nan	20.6000 ± 7.3103	0.0046 ± 0.0016	1300.0000 ± 458.2576	100.0%	5/8
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0.5$)	0.00522 ± 0.0028	1603.4000 ± 800.9541	100.0%	0.0010 ± 0.0000	19.3000 ± 6.9289	0.0044 ± 0.0025	1300.0000 ± 640.3124	100.0%	5/8
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0.75$)	0.00918 ± 0.0044	2808.2000 ± 1323.9158	100.0%	nan ± nan	20.5000 ± 10.0722	0.0064 ± 0.0021	2000.0000 ± 894.4272	100.0%	5/7
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 0.9$)	0.01082 ± 0.0053	3519.4000 ± 1687.8762	100.0%	0.0010 ± 0.0000	39.8000 ± 28.1169	0.0083 ± 0.0062	2700.0000 ± 1951.9221	100.0%	5/7
Q-learning ($\rho = \frac{1}{n(x,u)^{(0.7)}}, \epsilon = 1$)	0.01248 ± 0.0133	3919.2000 ± 4109.3820	100.0%	nan ± nan	18.5000 ± 11.5866	0.0052 ± 0.0022	1700.0000 ± 640.3124	100.0%	5/11

Table 31. Comparing performance of Q-learning as we change the learning rate ρ 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 ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.29275 ± 0.0132	84080.9000 ± 1846.8434	0.0%	0.0057 ± 0.0013	1967.5000 ± 306.9307	0.1255 ± 0.0115	36400.0000 ± 3006.6593	100.0%	18/20
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.26636 ± 0.0149	79924.8000 ± 2140.9895	0.0%	0.0065 ± 0.0018	1879.2000 ± 323.0742	0.1033 ± 0.0160	30250.0000 ± 4264.6805	80.0%	18/114
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.27654 ± 0.0171	80032.2000 ± 3053.1595	0.0%	0.0061 ± 0.0009	1855.3000 ± 164.8211	0.0892 ± 0.0526	26500.0000 ± 15394.8043	80.0%	18/23
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.27131 ± 0.0086	81149.9000 ± 2000.3920	0.0%	0.0056 ± 0.0011	1722.4000 ± 188.7470	0.2524 ± 0.0000	76000.0000 ± 0.0000	10.0%	18/24
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.36719 ± 0.0107	111279.2000 ± 292.4270	0.0%	0.0109 ± 0.0014	2919.9000 ± 214.8583	0.1260 ± 0.0040	37800.0000 ± 748.3315	100.0%	18/20
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.34508 ± 0.0084	102965.2000 ± 626.5941	0.0%	0.0092 ± 0.0010	2662.1000 ± 307.7208	0.1028 ± 0.0103	31200.0000 ± 3370.4599	100.0%	18/21
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	0.32791 ± 0.0084	99398.8000 ± 820.5098	0.0%	0.0085 ± 0.0011	2554.2000 ± 161.4861	0.1068 ± 0.0389	32800.0000 ± 12335.3152	100.0%	18/21
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	0.32642 ± 0.0084	99296.9000 ± 859.8602	0.0%	0.0083 ± 0.0009	2534.2000 ± 228.4088	0.2041 ± 0.0803	63666.6667 ± 24870.7771	60.0%	18/20
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.52335 ± 0.0127	163961.0000 ± 616.8108	0.0%	0.0119 ± 0.0015	3636.8000 ± 338.8064	0.1414 ± 0.0087	44600.0000 ± 3527.0384	100.0%	18/20
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.51433 ± 0.0258	155895.7000 ± 828.0388	0.0%	0.0121 ± 0.0020	3510.1000 ± 294.2127	0.1107 ± 0.0083	33800.0000 ± 1400.0000	100.0%	18/20
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	0.48894 ± 0.0179	151876.9000 ± 636.4334	0.0%	0.0120 ± 0.0028	3323.0000 ± 342.3609	0.1198 ± 0.0408	36700.0000 ± 12329.2336	100.0%	18/20
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.48487 ± 0.0111	152445.2000 ± 1059.5956	0.0%	0.0110 ± 0.0017	3386.7000 ± 394.3341	0.2083 ± 0.0834	64000.0000 ± 24236.9258	70.0%	18/23
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	0.99484 ± 0.0309	328077.4000 ± 2109.0571	0.0%	0.0139 ± 0.0030	4274.7000 ± 831.4579	0.1716 ± 0.0108	56300.0000 ± 3407.3450	100.0%	18/20
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	0.95465 ± 0.0142	323425.1000 ± 3302.4849	0.0%	0.0126 ± 0.0024	4013.7000 ± 829.6018	0.1388 ± 0.0181	47000.0000 ± 4898.9795	100.0%	18/20
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	0.97377 ± 0.0141	324972.4000 ± 2668.7384	0.0%	0.0109 ± 0.0033	3852.1000 ± 1009.1040	0.1309 ± 0.0300	44100.0000 ± 10270.8325	100.0%	18/19
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	1.02503 ± 0.0235	340748.4000 ± 3016.0495	0.0%	0.0111 ± 0.0040	3589.9000 ± 1256.2723	0.6227 ± 0.2873	207714.2857 ± 94152.9368	70.0%	18/20
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	2.53626 ± 0.0693	864966.5000 ± 7960.5387	0.0%	0.0112 ± 0.0049	3411.1000 ± 1898.8123	0.2336 ± 0.0217	79900.0000 ± 6394.5289	100.0%	18/18
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	2.53804 ± 0.0382	874687.6000 ± 9397.7727	0.0%	0.0092 ± 0.0053	3151.2000 ± 1792.1591	0.1904 ± 0.0211	64100.0000 ± 7133.7227	100.0%	18/20
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	2.73888 ± 0.0292	946961.9000 ± 7960.9170	0.0%	0.0095 ± 0.0040	3250.1000 ± 1394.9028	0.1693 ± 0.0240	58000.0000 ± 7797.4355	100.0%	18/20
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	3.12116 ± 0.0392	1081335.3000 ± 14219.5629	0.0%	0.0075 ± 0.0037	2562.5000 ± 1289.2223	1.3253 ± 0.8370	459750.0000 ± 291632.7957	80.0%	18/18
Q-learning ($\rho = 0.5, \epsilon = 1$)	11.29307 ± 0.1398	3939097.5000 ± 45977.1953	0.0%	0.0039 ± 0.0029	1233.4000 ± 967.1499	0.3111 ± 0.0539	109800.0000 ± 17244.1294	100.0%	18/19
Q-learning ($\rho = 0.7, \epsilon = 1$)	11.24880 ± 0.1818	3924276.5000 ± 39537.0995	0.0%	0.0033 ± 0.0018	1063.5000 ± 753.2162	0.3056 ± 0.0663	106900.0000 ± 22858.0402	100.0%	18/19
Q-learning ($\rho = 0.9, \epsilon = 1$)	11.29333 ± 0.1449	3936524.3000 ± 31734.1047	0.0%	0.0045 ± 0.0041	1477.4000 ± 1470.0947	0.2744 ± 0.0814	95800.0000 ± 28582.5121	100.0%	18/20
Q-learning ($\rho = 0.99, \epsilon = 1$)	11.28905 ± 0.1650	3953886.3000 ± 46600.8725	0.0%	0.0040 ± 0.0035	1235.4000 ± 1202.0360	5.3143 ± 2.6746	1865625.0000 ± 946252.0723	80.0%	18/19
Async Value Iteration	0.13340 ± 0.0056	212072.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	18/18
Value Iteration	0.18218 ± 0.0049	285648.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	18/20

Table 32. Comparing performance of Q-learning as we change the learning rate ρ 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 ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.27947 ± 0.0141	81180.4000 ± 168.4661	0.0%	0.0052 ± 0.0010	1574.9000 ± 56.2040	0.1184 ± 0.0102	34200.0000 ± 600.0000	100.0%	18/18
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.26669 ± 0.0288	73372.5000 ± 210.4933	0.0%	0.0059 ± 0.0023	1657.2000 ± 103.7630	0.0918 ± 0.0149	25200.0000 ± 400.0000	100.0%	18/18
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.22992 ± 0.0057	69749.8000 ± 575.6870	0.0%	0.0054 ± 0.0007	1579.3000 ± 62.5620	0.0684 ± 0.0039	20400.0000 ± 800.0000	100.0%	18/20
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.23145 ± 0.0126	70374.2000 ± 2204.0658	0.0%	0.0050 ± 0.0006	1554.4000 ± 90.5044	0.0625 ± 0.0033	19333.3333 ± 816.4966	90.0%	18/20
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.35672 ± 0.0059	108971.1000 ± 529.4137	0.0%	0.0089 ± 0.0021	2812.9000 ± 446.2606	0.1235 ± 0.0045	37100.0000 ± 300.0000	100.0%	18/19
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.33175 ± 0.0136	100219.0000 ± 305.3116	0.0%	0.0082 ± 0.0012	2548.6000 ± 203.3791	0.0924 ± 0.0032	27900.0000 ± 538.5165	100.0%	18/18
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	0.31436 ± 0.0079	96314.3000 ± 329.4295	0.0%	0.0085 ± 0.0016	2585.8000 ± 365.8658	0.0732 ± 0.0045	22100.0000 ± 300.0000	100.0%	18/18
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	0.31680 ± 0.0073	96356.5000 ± 611.5939	0.0%	0.0082 ± 0.0014	2614.5000 ± 317.4631	0.0685 ± 0.0032	21000.0000 ± 447.2136	100.0%	18/18
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.51660 ± 0.0110	160177.5000 ± 909.7031	0.0%	0.0107 ± 0.0018	3508.9000 ± 342.5149	0.1372 ± 0.0054	42200.0000 ± 400.0000	100.0%	18/18
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.48789 ± 0.0084	154458.7000 ± 1116.4002	0.0%	0.0111 ± 0.0017	3425.3000 ± 423.0629	0.1020 ± 0.0025	32100.0000 ± 300.0000	100.0%	18/18
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	0.48580 ± 0.0073	153836.3000 ± 1518.3678	0.0%	0.0114 ± 0.0016	3474.4000 ± 328.8663	0.0849 ± 0.0051	26900.0000 ± 830.6624	100.0%	18/18
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.52147 ± 0.0247	160492.7000 ± 1837.0062	0.0%	0.0099 ± 0.0017	3181.8000 ± 402.3264	0.0799 ± 0.0025	25400.0000 ± 489.8979	100.0%	18/19
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	1.04260 ± 0.0694	341004.1000 ± 7067.5569	0.0%	0.0151 ± 0.0049	4830.3000 ± 1113.3672	0.1628 ± 0.0157	53300.0000 ± 1004.9876	100.0%	18/18
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	1.09828 ± 0.0847	352123.5000 ± 5513.8133	0.0%	0.0143 ± 0.0035	4408.8000 ± 628.1254	0.1307 ± 0.0103	42100.0000 ± 943.3981	100.0%	18/20
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	1.30328 ± 0.0317	393504.4000 ± 8608.0014	0.0%	0.0140 ± 0.0031	4230.1000 ± 889.7672	0.1199 ± 0.0064	36000.0000 ± 1095.4451	100.0%	18/18
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	1.34013 ± 0.0416	429710.4000 ± 7278.7790	0.0%	0.0121 ± 0.0028	3983.6000 ± 953.3367	0.1077 ± 0.0071	34500.0000 ± 1627.8821	100.0%	18/19
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	0.99946 ± 0.1518	334382.4000 ± 45421.2774	100.0%	0.0112 ± 0.0054	3520.9000 ± 1323.5073	0.2142 ± 0.0158	70500.0000 ± 2801.7851	100.0%	18/18
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	0.77184 ± 0.1358	263606.7000 ± 47568.5157	100.0%	0.0116 ± 0.0043	3979.0000 ± 1540.8138	0.1631 ± 0.0057	54700.0000 ± 1552.4175	100.0%	18/18
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	0.65356 ± 0.0839	209086.1000 ± 18344.3354	100.0%	0.0106 ± 0.0042	3591.6000 ± 1285.8875	0.1468 ± 0.0099	48400.0000 ± 2154.0659	100.0%	18/18
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	0.64452 ± 0.1320	179820.8000 ± 16648.6672	100.0%	0.0145 ± 0.0069	4169.9000 ± 1901.6330	0.1643 ± 0.0460	46400.0000 ± 1907.8784	100.0%	18/18
Q-learning ($\rho = 0.5, \epsilon = 1$)	0.74025 ± 0.1282	213183.3000 ± 14395.8089	100.0%	0.0061 ± 0.0060	1798.9000 ± 1729.0229	0.3303 ± 0.0757	95000.0000 ± 7771.7437	100.0%	18/18
Q-learning ($\rho = 0.7, \epsilon = 1$)	0.60477 ± 0.1074	167509.4000 ± 12420.1262	100.0%	0.0112 ± 0.0110	1994.8000 ± 1630.5474	0.2904 ± 0.0727	77600.0000 ± 7459.2225	100.0%	18/18
Q-learning ($\rho = 0.9, \epsilon = 1$)	0.49257 ± 0.0890	146319.4000 ± 12333.1839	100.0%	0.0039 ± 0.0027	1297.3000 ± 858.0210	0.2339 ± 0.0261	71100.0000 ± 4805.2055	100.0%	18/18
Q-learning ($\rho = 0.99, \epsilon = 1$)	0.50919 ± 0.1118	143606.5000 ± 18301.5489	100.0%	0.0050 ± 0.0034	1497.0000 ± 1140.6033	0.2263 ± 0.0370	64600.0000 ± 3666.0606	100.0%	18/18
Async Value Iteration	0.12769 ± 0.0224	177448.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	18/18
Value Iteration	0.16512 ± 0.0425	225056.0000 ± 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 ρ 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 ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.38042 ± 0.0107	117489.8000 ± 85.4222	0.0%	0.0032 ± 0.0009	991.8000 ± 251.1975	0.1427 ± 0.0110	44200.0000 ± 2712.9320	100.0%	30/31
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.35820 ± 0.0100	110759.3000 ± 1304.7708	0.0%	0.0032 ± 0.0009	974.8000 ± 185.1523	0.1083 ± 0.0107	32700.0000 ± 2647.6405	100.0%	30/32
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.36663 ± 0.0224	109215.5000 ± 1517.1100	0.0%	0.0033 ± 0.0011	999.0000 ± 274.0759	0.0868 ± 0.0092	26200.0000 ± 2521.9040	100.0%	30/808
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.35544 ± 0.0085	109650.5000 ± 1558.8350	0.0%	0.0027 ± 0.0015	860.2000 ± 366.6846	0.1494 ± 0.0523	44714.2857 ± 14771.7325	70.0%	30/36
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.47809 ± 0.0085	151503.0000 ± 254.3887	0.0%	0.0037 ± 0.0013	1166.5000 ± 400.4625	0.1958 ± 0.0295	62200.0000 ± 8863.4079	100.0%	30/32
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.45357 ± 0.0066	144102.2000 ± 364.9177	0.0%	0.0037 ± 0.0009	1249.3000 ± 324.4814	0.1913 ± 0.1033	60222.2222 ± 32054.7371	90.0%	30/34
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	0.45130 ± 0.0140	140627.8000 ± 412.1582	0.0%	0.0039 ± 0.0017	1143.6000 ± 478.0871	0.1857 ± 0.0937	58500.0000 ± 30500.0000	20.0%	30/31
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	0.44438 ± 0.0087	139857.0000 ± 391.0667	0.0%	0.0041 ± 0.0016	1236.3000 ± 444.9135	nan ± nan	nan ± nan	0.0%	30/33
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.69795 ± 0.0119	224149.9000 ± 884.7431	0.0%	0.0058 ± 0.0026	1882.9000 ± 892.0393	0.2534 ± 0.0829	81500.0000 ± 25831.1827	100.0%	30/34
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.67333 ± 0.0114	217890.6000 ± 1170.2144	0.0%	0.0044 ± 0.0016	1380.1000 ± 508.8783	0.3178 ± 0.1220	102625.0000 ± 40493.6338	80.0%	30/31
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	0.66646 ± 0.0091	215497.8000 ± 720.2024	0.0%	0.0067 ± 0.0031	2142.5000 ± 874.4046	0.1926 ± 0.0761	62500.0000 ± 24500.0000	20.0%	30/32
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.66310 ± 0.0108	214407.5000 ± 781.2234	0.0%	0.0043 ± 0.0016	1384.1000 ± 524.9748	nan ± nan	nan ± nan	0.0%	30/32
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	1.42015 ± 0.0129	471036.8000 ± 3448.4362	0.0%	0.0066 ± 0.0029	2242.7000 ± 984.2487	0.6290 ± 0.3200	208444.4444 ± 105667.3093	90.0%	30/33
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	1.40379 ± 0.0162	467047.8000 ± 2444.8073	0.0%	0.0058 ± 0.0026	1966.1000 ± 777.4724	0.7169 ± 0.3196	241125.0000 ± 107499.3459	80.0%	30/32
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	1.41994 ± 0.0211	471120.3000 ± 3976.5778	0.0%	0.0067 ± 0.0049	2320.0000 ± 1454.0864	1.0996 ± 0.0689	362666.6667 ± 21929.1789	30.0%	30/1773
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	1.44096 ± 0.0122	477468.3000 ± 1642.4698	0.0%	0.0073 ± 0.0032	2398.5000 ± 1099.7910	nan ± nan	nan ± nan	0.0%	30/32
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	3.80891 ± 0.0409	1280614.8000 ± 11464.5819	0.0%	0.0062 ± 0.0024	1949.9000 ± 814.3826	0.6808 ± 0.4437	229800.0000 ± 147197.6902	100.0%	30/32
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	3.83213 ± 0.0397	1291548.1000 ± 10933.6311	0.0%	0.0094 ± 0.0064	3205.1000 ± 2052.1897	0.7794 ± 0.3948	263400.0000 ± 130889.4190	100.0%	30/31
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	3.87834 ± 0.0649	1309331.5000 ± 20062.2936	0.0%	0.0067 ± 0.0023	2035.4000 ± 814.8637	1.5721 ± 1.0353	531444.4444 ± 351328.9409	90.0%	30/32
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	4.11874 ± 0.0570	1381097.7000 ± 12916.8569	0.0%	0.0072 ± 0.0035	2372.5000 ± 1232.2879	nan ± nan	nan ± nan	0.0%	30/32
Q-learning ($\rho = 0.5, \epsilon = 1$)	15.96155 ± 0.1005	5434790.2000 ± 33827.7444	0.0%	0.0072 ± 0.0052	2469.3000 ± 1684.1995	1.1400 ± 0.4229	387400.0000 ± 142776.1885	100.0%	30/31
Q-learning ($\rho = 0.7, \epsilon = 1$)	15.98997 ± 0.2010	5452153.0000 ± 49448.4142	0.0%	0.0040 ± 0.0041	1295.0000 ± 1398.0005	0.6707 ± 0.3531	230600.0000 ± 123099.3095	100.0%	30/31
Q-learning ($\rho = 0.9, \epsilon = 1$)	15.98565 ± 0.1534	5436481.1000 ± 48745.8184	0.0%	0.0055 ± 0.0041	1862.2000 ± 1310.0016	2.9607 ± 2.2130	1009888.8889 ± 751642.7113	90.0%	30/32
Q-learning ($\rho = 0.99, \epsilon = 1$)	15.98029 ± 0.1735	5427919.6000 ± 67890.6968	0.0%	0.0052 ± 0.0022	1667.9000 ± 755.2251	2.6790 ± 0.0000	897000.0000 ± 0.0000	10.0%	30/32
Async Value Iteration	0.07369 ± 0.0033	103156.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	30/32
Value Iteration	0.13184 ± 0.0026	186184.0000 ± 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 ρ in stochastic problem with $\gamma = 0.999$ (Problem 2). Two versions of value iteration are also appended for contrast

Algorithm (Problem 2) ($\gamma = 0.999$)	Runtime (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.37391 ± 0.0096	115387.7000 ± 90.8483	0.0%	0.0028 ± 0.0004	888.1000 ± 54.9371	0.1367 ± 0.0067	41300.0000 ± 458.2576	100.0%	30/30
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.34651 ± 0.0059	107608.8000 ± 142.4449	0.0%	0.0031 ± 0.0007	834.1000 ± 106.5143	0.0974 ± 0.0043	29900.0000 ± 300.0000	100.0%	30/30
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.33434 ± 0.0109	103532.0000 ± 120.4691	0.0%	0.0028 ± 0.0004	873.6000 ± 54.8675	0.0738 ± 0.0040	23000.0000 ± 0.0000	100.0%	30/30
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.33135 ± 0.0113	102770.4000 ± 326.7722	0.0%	0.0029 ± 0.0006	887.7000 ± 57.4300	0.0688 ± 0.0033	21000.0000 ± 0.0000	100.0%	30/30
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.47867 ± 0.0104	149327.3000 ± 362.5518	0.0%	0.0038 ± 0.0014	1156.8000 ± 327.9463	0.1473 ± 0.0051	45800.0000 ± 400.0000	100.0%	30/30
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.44759 ± 0.0144	141408.5000 ± 298.7170	0.0%	0.0037 ± 0.0014	1208.3000 ± 417.0645	0.1065 ± 0.0064	33200.0000 ± 400.0000	100.0%	30/30
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	0.43571 ± 0.0110	137861.8000 ± 351.7766	0.0%	0.0044 ± 0.0019	1359.7000 ± 490.6096	0.0837 ± 0.0032	26400.0000 ± 489.8979	100.0%	30/30
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	0.43773 ± 0.0085	138153.3000 ± 424.4967	0.0%	0.0041 ± 0.0017	1203.2000 ± 411.3796	0.0790 ± 0.0034	24300.0000 ± 458.2576	100.0%	30/31
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.68519 ± 0.0102	220500.2000 ± 1160.6733	0.0%	0.0057 ± 0.0018	1857.0000 ± 580.8521	0.1651 ± 0.0067	53100.0000 ± 830.6624	100.0%	30/30
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.66315 ± 0.0141	214552.2000 ± 903.8086	0.0%	0.0039 ± 0.0023	1363.7000 ± 753.4771	0.1220 ± 0.0048	39500.0000 ± 670.8204	100.0%	30/30
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	0.66222 ± 0.0131	214508.6000 ± 1366.8492	0.0%	0.0052 ± 0.0023	1626.4000 ± 695.1079	0.1015 ± 0.0036	32300.0000 ± 458.2576	100.0%	30/30
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.68010 ± 0.0160	218352.0000 ± 1178.6356	0.0%	0.0056 ± 0.0025	1677.8000 ± 630.3829	0.0903 ± 0.0068	29500.0000 ± 806.2258	100.0%	30/32
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	1.42425 ± 0.0689	469466.8000 ± 23214.5110	20.0%	0.0090 ± 0.0036	2965.7000 ± 1074.8678	0.2028 ± 0.0089	67500.0000 ± 1284.5233	100.0%	30/30
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	1.02437 ± 0.2339	340720.1000 ± 77353.3107	80.0%	0.0083 ± 0.0036	2751.5000 ± 1179.5511	0.1522 ± 0.0068	51000.0000 ± 1183.2160	100.0%	30/30
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	0.83226 ± 0.3018	277461.3000 ± 101749.3286	90.0%	0.0047 ± 0.0017	1717.8000 ± 637.6949	0.1321 ± 0.0074	43800.0000 ± 1720.4651	100.0%	30/32
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	0.73714 ± 0.2768	245679.5000 ± 92101.5595	100.0%	0.0066 ± 0.0025	2146.4000 ± 752.7289	0.1198 ± 0.0044	40500.0000 ± 1204.1595	100.0%	30/30
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	0.45782 ± 0.0498	155604.3000 ± 15067.6532	100.0%	0.0057 ± 0.0036	2102.6000 ± 1253.4483	0.2607 ± 0.0145	88400.0000 ± 5180.7335	100.0%	30/30
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	0.37586 ± 0.0407	128528.2000 ± 15599.7919	100.0%	0.0106 ± 0.0060	3165.5000 ± 2059.8085	0.2027 ± 0.0061	69100.0000 ± 1813.8357	100.0%	30/30
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	0.30046 ± 0.0259	102929.4000 ± 9634.7885	100.0%	0.0074 ± 0.0038	2505.7000 ± 1270.0775	0.1719 ± 0.0088	58700.0000 ± 3001.6662	100.0%	30/30
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	0.26071 ± 0.0244	88445.9000 ± 7099.9132	100.0%	0.0086 ± 0.0046	2815.5000 ± 1537.1078	0.1616 ± 0.0116	54900.0000 ± 3419.0642	100.0%	30/30
Q-learning ($\rho = 0.5, \epsilon = 1$)	0.69021 ± 0.0701	236471.8000 ± 22304.3683	100.0%	0.0059 ± 0.0055	2177.4000 ± 1872.7837	0.5234 ± 0.0512	179600.0000 ± 17135.9272	100.0%	30/32
Q-learning ($\rho = 0.7, \epsilon = 1$)	0.57106 ± 0.0819	195922.6000 ± 25598.0802	100.0%	0.0066 ± 0.0044	2125.4000 ± 1381.2143	0.4280 ± 0.0609	149000.0000 ± 20327.3215	100.0%	30/30
Q-learning ($\rho = 0.9, \epsilon = 1$)	0.45826 ± 0.0505	158138.2000 ± 18294.6458	100.0%	0.0086 ± 0.0075	2860.8000 ± 2519.9723	0.3406 ± 0.0446	117600.0000 ± 16088.5052	100.0%	30/31
Q-learning ($\rho = 0.99, \epsilon = 1$)	0.41992 ± 0.0421	144132.2000 ± 16043.6442	100.0%	0.0046 ± 0.0025	1594.8000 ± 779.0513	0.3483 ± 0.0333	119100.0000 ± 11819.0524	100.0%	30/30
Async Value Iteration	0.05673 ± 0.0018	80512.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	30/32
Value Iteration	0.10499 ± 0.0041	148444.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	30/31

Table 35. Comparing performance of Q-learning as we change the learning rate ρ in stochastic problem with $\gamma = 0.99$ (Problem 3). Two versions of value iteration are also appended for contrast

Algorithm (Problem 3) ($\gamma = 0.99$)	Runtime (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.91569 ± 0.0220	305083.6000 ± 829.1153	0.0%	0.0029 ± 0.0011	923.4000 ± 383.4085	0.1681 ± 0.0097	56000.0000 ± 2683.2816	100.0%	91/93
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.89906 ± 0.0126	298119.3000 ± 1676.7089	0.0%	0.0039 ± 0.0018	1324.7000 ± 612.7228	0.1220 ± 0.0157	40800.0000 ± 4995.9984	100.0%	91/99
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.90295 ± 0.0144	300908.5000 ± 2148.4727	0.0%	0.0029 ± 0.0015	1047.9000 ± 389.2044	0.1653 ± 0.1309	55400.0000 ± 41924.2173	100.0%	91/2414
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.92180 ± 0.0105	306487.1000 ± 1610.2135	0.0%	0.0031 ± 0.0011	1074.4000 ± 285.4968	0.2525 ± 0.2590	85333.3333 ± 87598.8331	60.0%	94/2698
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	1.19079 ± 0.0177	404577.4000 ± 457.2989	0.0%	0.0039 ± 0.0015	1315.9000 ± 533.6925	0.2229 ± 0.0462	76500.0000 ± 15628.4996	100.0%	91/96
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	1.17456 ± 0.0189	397570.2000 ± 551.1620	0.0%	0.0040 ± 0.0020	1403.1000 ± 660.9223	0.1681 ± 0.0394	57000.0000 ± 13623.5091	100.0%	91/95
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	1.19039 ± 0.0189	401419.1000 ± 1175.2747	0.0%	0.0039 ± 0.0017	1393.3000 ± 646.9668	0.1491 ± 0.0725	50400.0000 ± 24658.4671	100.0%	91/385
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	1.20617 ± 0.0179	409324.8000 ± 1042.5569	0.0%	0.0035 ± 0.0019	1135.8000 ± 584.4448	0.3942 ± 0.4047	134111.1111 ± 138336.5015	90.0%	91/1565
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	1.75548 ± 0.0205	605848.2000 ± 1113.6285	0.0%	0.0065 ± 0.0040	2428.8000 ± 1452.7156	0.2190 ± 0.0334	75800.0000 ± 11830.4691	100.0%	91/96
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	1.74981 ± 0.0252	599233.0000 ± 1594.7029	0.0%	0.0052 ± 0.0024	1787.9000 ± 638.0621	0.1783 ± 0.0362	62500.0000 ± 12595.6342	100.0%	91/97
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	1.75534 ± 0.0541	603593.2000 ± 811.8044	0.0%	0.0037 ± 0.0018	1429.8000 ± 609.1137	0.1566 ± 0.0432	54600.0000 ± 15252.5408	100.0%	91/100002
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	1.79159 ± 0.0499	610444.0000 ± 1129.0487	0.0%	0.0061 ± 0.0042	2183.9000 ± 1382.0394	0.7486 ± 0.5684	259125.0000 ± 196897.4590	80.0%	91/100002
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	3.48230 ± 0.0493	1220420.2000 ± 2871.4701	0.0%	0.0081 ± 0.0047	2880.0000 ± 1585.0276	0.2669 ± 0.0241	95500.0000 ± 8766.4132	100.0%	91/98
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	3.45873 ± 0.0280	1221777.0000 ± 4075.2041	0.0%	0.0087 ± 0.0052	3206.5000 ± 1789.1007	0.2086 ± 0.0217	74700.0000 ± 8426.7431	100.0%	91/93
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	3.49219 ± 0.0454	1229891.2000 ± 3135.3444	0.0%	0.0127 ± 0.0059	4464.7000 ± 2112.3342	0.2226 ± 0.0626	79900.0000 ± 22451.9487	100.0%	91/1522
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	3.50311 ± 0.0482	1236475.6000 ± 5450.0068	0.0%	0.0076 ± 0.0042	2696.8000 ± 1480.7171	1.4869 ± 0.7039	528571.4286 ± 250881.2224	70.0%	91/7122
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	8.86108 ± 0.0917	3163606.7000 ± 19721.0124	0.0%	0.0212 ± 0.0126	7924.6000 ± 4719.8429	0.3457 ± 0.0404	124300.0000 ± 12884.4868	100.0%	91/96
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	8.89460 ± 0.0740	3170117.5000 ± 9883.4341	0.0%	0.0238 ± 0.0136	8591.9000 ± 4700.7429	0.2869 ± 0.0421	103900.0000 ± 15616.9779	100.0%	91/94
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	8.88675 ± 0.0945	3177721.4000 ± 26601.0311	0.0%	0.0165 ± 0.0101	5927.1000 ± 3466.5689	0.3222 ± 0.0949	115800.0000 ± 34617.3367	100.0%	91/676
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	8.88070 ± 0.1104	3182320.0000 ± 19246.8723	0.0%	0.0125 ± 0.0085	4576.2000 ± 3065.0236	2.8284 ± 1.9855	1008500.0000 ± 707036.8095	100.0%	91/1524
Q-learning ($\rho = 0.5, \epsilon = 1$)	24.27070 ± 0.1993	8840989.5000 ± 9751.2095	0.0%	0.1485 ± 0.1418	53959.9000 ± 50084.4462	1.5655 ± 0.3155	574100.0000 ± 114425.0410	100.0%	91/95
Q-learning ($\rho = 0.7, \epsilon = 1$)	2387.57080 ± 7088.5490	8839172.9000 ± 14993.4189	0.0%	0.0984 ± 0.0821	34859.8000 ± 28505.7054	1.1087 ± 0.3218	400100.0000 ± 110732.5155	100.0%	91/95
Q-learning ($\rho = 0.9, \epsilon = 1$)	24.98004 ± 0.3991	8832810.2000 ± 14496.8004	0.0%	0.1154 ± 0.0731	41696.8000 ± 26378.4637	1.0684 ± 0.3453	372400.0000 ± 103835.6394	100.0%	91/971
Q-learning ($\rho = 0.99, \epsilon = 1$)	24.84772 ± 0.1943	8831593.4000 ± 13980.3939	0.0%	0.1373 ± 0.1126	49139.0000 ± 38936.7101	4.4245 ± 4.4615	1577500.0000 ± 1588784.7085	100.0%	91/1319
Async Value Iteration	0.04747 ± 0.0019	46580.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	91/98
Value Iteration	0.08371 ± 0.0031	78364.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	91/97

Table 36. Comparing performance of Q-learning as we change the learning rate ρ in stochastic problem with $\gamma = 0.999$ (Problem 3). Two versions of value iteration are also appended for contrast

Algorithm (Problem 3) ($\gamma = 0.999$)	Runtime (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.92399 ± 0.0137	300275.0000 ± 43.3428	0.0%	0.0023 ± 0.0007	735.3000 ± 57.0299	0.1622 ± 0.0085	53000.0000 ± 0.0000	100.0%	91/93
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.89158 ± 0.0111	292385.3000 ± 1129.4207	0.0%	0.0026 ± 0.0009	754.6000 ± 221.6552	0.1127 ± 0.0029	37700.0000 ± 640.3124	100.0%	91/93
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.88467 ± 0.0109	288919.2000 ± 1479.9050	0.0%	0.0023 ± 0.0007	742.3000 ± 175.9023	0.0924 ± 0.0063	30400.0000 ± 1624.8077	100.0%	91/94
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.88749 ± 0.0165	290238.9000 ± 1683.4562	0.0%	0.0022 ± 0.0006	735.1000 ± 176.5794	0.1073 ± 0.0300	35300.0000 ± 9445.1046	100.0%	93/96
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.33719 ± 0.0355	110853.4000 ± 11286.0892	100.0%	0.0050 ± 0.0022	1655.0000 ± 690.3288	0.1819 ± 0.0089	59700.0000 ± 1268.8578	100.0%	91/91
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.35709 ± 0.1380	117570.6000 ± 45850.4824	100.0%	0.0039 ± 0.0018	1230.1000 ± 543.7403	0.1335 ± 0.0073	44700.0000 ± 3034.7982	100.0%	91/93
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	0.90321 ± 0.3797	303684.4000 ± 128084.9031	40.0%	0.0028 ± 0.0009	1066.3000 ± 372.9850	0.1054 ± 0.0070	35600.0000 ± 1280.6248	100.0%	91/93
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	1.13980 ± 0.0080	386642.8000 ± 535.2334	0.0%	0.0042 ± 0.0022	1452.9000 ± 644.1227	0.1645 ± 0.0830	56800.0000 ± 28385.9120	100.0%	91/91
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.25907 ± 0.0275	89297.4000 ± 9104.6975	100.0%	0.0057 ± 0.0035	2073.1000 ± 1219.3274	0.2033 ± 0.0078	69800.0000 ± 1166.1904	100.0%	91/93
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.22204 ± 0.0355	75587.1000 ± 12552.5460	100.0%	0.0081 ± 0.0049	2715.7000 ± 1714.5618	0.1504 ± 0.0072	51500.0000 ± 1360.1471	100.0%	91/92
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	1.19841 ± 0.4680	416605.8000 ± 162564.7256	60.0%	0.0049 ± 0.0024	1690.9000 ± 744.8889	0.1329 ± 0.0206	47100.0000 ± 7462.5733	100.0%	91/91
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	1.69167 ± 0.0139	587830.2000 ± 1091.2304	0.0%	0.0069 ± 0.0046	2479.2000 ± 1649.8408	0.1796 ± 0.0400	63000.0000 ± 13718.6005	100.0%	91/91
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	0.33701 ± 0.0470	118399.0000 ± 16698.0658	100.0%	0.0080 ± 0.0039	2672.7000 ± 1375.4718	0.2548 ± 0.0152	90100.0000 ± 4085.3396	100.0%	91/91
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	0.34945 ± 0.1720	123187.6000 ± 58510.0831	100.0%	0.0107 ± 0.0040	3766.2000 ± 1390.1361	0.1912 ± 0.0109	68300.0000 ± 1791.6473	100.0%	91/91
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	0.82516 ± 0.3671	289907.2000 ± 127240.3462	100.0%	0.0099 ± 0.0060	3416.6000 ± 1975.4050	0.1838 ± 0.0291	65700.0000 ± 9022.7490	100.0%	91/93
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	2.70835 ± 1.0825	955920.9000 ± 380635.6415	40.0%	0.0080 ± 0.0052	2892.2000 ± 1626.8171	0.1866 ± 0.0403	66700.0000 ± 14262.1878	100.0%	91/93
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	0.44541 ± 0.1079	159110.3000 ± 39435.1220	100.0%	0.0135 ± 0.0058	4952.4000 ± 1993.6882	0.3330 ± 0.0210	119100.0000 ± 6300.0000	100.0%	91/93
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	0.42875 ± 0.1005	153042.4000 ± 33703.6039	100.0%	0.0179 ± 0.0095	6541.5000 ± 3600.3277	0.2903 ± 0.0347	103800.0000 ± 11822.0134	100.0%	91/91
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	0.83630 ± 0.6067	296035.1000 ± 215822.0164	100.0%	0.0203 ± 0.0073	7183.3000 ± 2602.6955	0.2693 ± 0.0450	96000.0000 ± 16685.3229	100.0%	91/91
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	2.89879 ± 1.3965	1035424.9000 ± 501427.7863	100.0%	0.0187 ± 0.0090	6735.3000 ± 3253.0974	0.2911 ± 0.0999	105500.0000 ± 36951.9959	100.0%	91/93
Q-learning ($\rho = 0.5, \epsilon = 1$)	1.85509 ± 0.4075	682548.3000 ± 148915.5958	100.0%	0.0859 ± 0.0763	32055.1000 ± 28241.9606	1.3763 ± 0.1824	505900.0000 ± 67291.0841	100.0%	91/91
Q-learning ($\rho = 0.7, \epsilon = 1$)	1.66867 ± 0.2670	614939.0000 ± 99674.4243	100.0%	0.1237 ± 0.1109	45932.1000 ± 40450.8210	1.2381 ± 0.2881	455900.0000 ± 106082.4679	100.0%	91/93
Q-learning ($\rho = 0.9, \epsilon = 1$)	1.46615 ± 0.3284	540734.0000 ± 124010.2209	100.0%	0.0929 ± 0.0926	34287.8000 ± 34755.0378	1.0646 ± 0.2174	392100.0000 ± 82443.2532	100.0%	91/93
Q-learning ($\rho = 0.99, \epsilon = 1$)	1.88713 ± 0.7659	692405.6000 ± 278884.3892	100.0%	0.1848 ± 0.2254	67403.0000 ± 82015.0070	1.0949 ± 0.4034	400700.0000 ± 147694.9898	100.0%	91/92
Async Value Iteration	0.03797 ± 0.0008	39456.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	91/92
Value Iteration	0.06558 ± 0.0019	66856.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	91/91

Table 37. Comparing performance of Q-learning as we change the learning rate ρ in stochastic problem with $\gamma = 0.99$ (Problem 4). Two versions of value iteration are also appended for contrast

Algorithm (Problem 4) ($\gamma = 0.99$)	Runtime (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.40503 ± 0.0090	122956.5000 ± 187.7159	0.0%	0.0060 ± 0.0011	2000.6000 ± 373.1660	0.1559 ± 0.0085	47400.0000 ± 1496.6630	100.0%	32/33
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.38064 ± 0.0079	115530.4000 ± 244.0034	0.0%	0.0060 ± 0.0009	1896.8000 ± 391.8856	0.1247 ± 0.0105	38400.0000 ± 3231.0989	100.0%	32/895
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.37470 ± 0.0158	113162.5000 ± 1334.7168	0.0%	0.0067 ± 0.0014	2067.6000 ± 565.3532	0.1120 ± 0.0233	33400.0000 ± 7031.3583	100.0%	32/34
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.37323 ± 0.0113	113029.0000 ± 838.3514	0.0%	0.0062 ± 0.0017	1876.1000 ± 400.3562	0.1980 ± 0.0647	59750.0000 ± 18646.3803	80.0%	32/36
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.50657 ± 0.0146	156758.3000 ± 319.9447	0.0%	0.0067 ± 0.0010	2040.9000 ± 284.5560	0.3191 ± 0.1254	98375.0000 ± 35790.8421	80.0%	32/35
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.48630 ± 0.0214	149218.7000 ± 394.3108	0.0%	0.0068 ± 0.0020	1993.0000 ± 406.1746	0.2626 ± 0.1072	77714.2857 ± 31033.2277	70.0%	32/35
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	0.47289 ± 0.0088	145750.9000 ± 426.4940	0.0%	0.0053 ± 0.0012	1763.5000 ± 407.4826	0.4269 ± 0.0000	133000.0000 ± 0.0000	10.0%	32/34
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	0.46922 ± 0.0170	144790.4000 ± 476.4599	0.0%	0.0056 ± 0.0012	1897.2000 ± 448.9837	nan ± nan	nan ± nan	0.0%	32/32
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.71706 ± 0.0191	226451.1000 ± 900.6267	0.0%	0.0088 ± 0.0024	2726.7000 ± 902.7455	0.5134 ± 0.1309	161000.0000 ± 39217.3431	40.0%	32/33
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.68687 ± 0.0193	219192.1000 ± 688.5299	0.0%	0.0084 ± 0.0024	2689.8000 ± 734.4059	0.5305 ± 0.0000	161000.0000 ± 0.0000	10.0%	32/33
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	0.68526 ± 0.0143	217004.3000 ± 510.3309	0.0%	0.0084 ± 0.0021	2612.2000 ± 380.0742	0.4435 ± 0.0000	141000.0000 ± 0.0000	10.0%	32/33
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.67931 ± 0.0150	216135.7000 ± 1124.4547	0.0%	0.0088 ± 0.0023	2848.4000 ± 819.3611	nan ± nan	nan ± nan	0.0%	32/34
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	1.38091 ± 0.0359	449401.1000 ± 3093.4826	0.0%	0.0117 ± 0.0041	3814.4000 ± 1079.4925	0.6122 ± 0.3599	202333.3333 ± 119471.9865	60.0%	32/35
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	1.36619 ± 0.0221	442885.5000 ± 3242.3715	0.0%	0.0126 ± 0.0042	4226.3000 ± 1147.5300	0.7837 ± 0.3400	254000.0000 ± 112000.0000	20.0%	32/34
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	1.35500 ± 0.0301	443110.8000 ± 4224.0447	0.0%	0.0114 ± 0.0021	3675.4000 ± 797.4487	0.1780 ± 0.0000	57000.0000 ± 0.0000	10.0%	32/33
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	1.40246 ± 0.0316	454316.7000 ± 2489.3804	0.0%	0.0109 ± 0.0040	3457.8000 ± 1071.8539	nan ± nan	nan ± nan	0.0%	32/35
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	3.63487 ± 0.0620	1208587.4000 ± 15351.1565	0.0%	0.0140 ± 0.0082	4539.4000 ± 2513.6557	0.8928 ± 0.4479	297571.4286 ± 145528.4529	70.0%	32/34
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	3.66430 ± 0.0541	1217489.8000 ± 9354.4082	0.0%	0.0084 ± 0.0038	2922.2000 ± 1057.0239	1.7074 ± 0.8188	561250.0000 ± 265643.9111	40.0%	32/35
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	3.81922 ± 0.0808	1267998.7000 ± 14328.1010	0.0%	0.0115 ± 0.0038	3860.5000 ± 1346.6014	nan ± nan	nan ± nan	0.0%	32/32
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	4.10351 ± 0.0541	1364489.6000 ± 12287.0927	0.0%	0.0177 ± 0.0080	6046.3000 ± 2464.7909	nan ± nan	nan ± nan	0.0%	32/34
Q-learning ($\rho = 0.5, \epsilon = 1$)	17.19990 ± 0.1706	5817783.5000 ± 40056.8863	0.0%	0.0063 ± 0.0041	2220.9000 ± 1426.2539	4.4260 ± 4.3751	1503222.2222 ± 1479816.9390	90.0%	32/35
Q-learning ($\rho = 0.7, \epsilon = 1$)	17.28202 ± 0.2256	5823556.6000 ± 48151.9840	0.0%	0.0092 ± 0.0064	3279.9000 ± 2093.3654	6.5277 ± 4.7949	2203285.7143 ± 1612919.4218	70.0%	32/34
Q-learning ($\rho = 0.9, \epsilon = 1$)	17.31619 ± 0.1418	5830502.8000 ± 52893.3075	0.0%	0.0100 ± 0.0071	3113.5000 ± 2549.6916	6.4427 ± 1.0813	2156500.0000 ± 392500.0000	20.0%	32/35
Q-learning ($\rho = 0.99, \epsilon = 1$)	17.25979 ± 0.2067	5853768.0000 ± 70104.9362	0.0%	0.0142 ± 0.0129	5047.7000 ± 4696.5863	nan ± nan	nan ± nan	0.0%	32/35
Async Value Iteration	0.10939 ± 0.0057	149072.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	32/35
Value Iteration	0.14271 ± 0.0047	191664.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	32/33

Table 38. Comparing performance of Q-learning as we change the learning rate ρ 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 ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.40398 ± 0.0179	121020.0000 ± 137.4831	0.0%	0.0060 ± 0.0012	1810.0000 ± 291.3884	0.1476 ± 0.0064	44400.0000 ± 663.3250	100.0%	32/32
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.37363 ± 0.0109	113288.0000 ± 71.1632	0.0%	0.0063 ± 0.0013	1862.1000 ± 329.1220	0.1065 ± 0.0050	32000.0000 ± 0.0000	100.0%	32/1130
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.36194 ± 0.0089	109375.8000 ± 190.3107	0.0%	0.0057 ± 0.0014	1732.7000 ± 261.4330	0.0793 ± 0.0060	24600.0000 ± 489.8979	100.0%	32/32
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.35485 ± 0.0136	108461.6000 ± 218.4038	0.0%	0.0066 ± 0.0016	1841.5000 ± 342.7907	0.0734 ± 0.0037	22000.0000 ± 0.0000	100.0%	32/32
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.49504 ± 0.0162	154348.3000 ± 448.5671	0.0%	0.0069 ± 0.0017	2180.9000 ± 425.9602	0.1603 ± 0.0099	49700.0000 ± 640.3124	100.0%	32/32
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.46898 ± 0.0135	146897.7000 ± 574.0019	0.0%	0.0068 ± 0.0014	1991.7000 ± 321.9975	0.1155 ± 0.0089	36000.0000 ± 447.2136	100.0%	32/32
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	0.46353 ± 0.0139	143821.4000 ± 481.2052	0.0%	0.0056 ± 0.0010	1825.6000 ± 398.4219	0.0894 ± 0.0042	28400.0000 ± 489.8979	100.0%	32/32
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	0.46397 ± 0.0113	144073.9000 ± 520.8655	0.0%	0.0057 ± 0.0018	1761.4000 ± 357.4555	0.0812 ± 0.0044	25700.0000 ± 458.2576	100.0%	32/32
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.69266 ± 0.0182	223784.9000 ± 1279.7536	0.0%	0.0084 ± 0.0032	2616.5000 ± 762.0247	0.1786 ± 0.0086	57800.0000 ± 748.3315	100.0%	32/32
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.69919 ± 0.0262	219532.7000 ± 1334.0115	0.0%	0.0074 ± 0.0026	2464.8000 ± 823.2115	0.1344 ± 0.0063	42900.0000 ± 538.5165	100.0%	32/33
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	0.69944 ± 0.0101	219236.7000 ± 1706.7261	0.0%	0.0076 ± 0.0014	2390.3000 ± 470.6778	0.1104 ± 0.0050	34600.0000 ± 800.0000	100.0%	32/32
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.69905 ± 0.0148	221584.3000 ± 627.8771	0.0%	0.0072 ± 0.0015	2298.4000 ± 471.4453	0.0957 ± 0.0058	31300.0000 ± 781.0250	100.0%	32/32
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	1.39298 ± 0.0279	451896.4000 ± 5107.7125	0.0%	0.0086 ± 0.0022	2908.8000 ± 904.4703	0.2212 ± 0.0092	72700.0000 ± 1268.8578	100.0%	32/32
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	1.44149 ± 0.0325	464625.9000 ± 3045.6726	0.0%	0.0107 ± 0.0037	3462.7000 ± 985.1840	0.1709 ± 0.0109	56300.0000 ± 1486.6069	100.0%	32/32
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	1.48087 ± 0.0261	479659.4000 ± 4032.2816	0.0%	0.0114 ± 0.0038	3664.8000 ± 1090.2903	0.1417 ± 0.0055	46900.0000 ± 700.0000	100.0%	32/32
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	1.51294 ± 0.0371	491389.0000 ± 3876.1084	0.0%	0.0110 ± 0.0033	3501.6000 ± 1062.4862	0.1314 ± 0.0069	43200.0000 ± 1939.0719	100.0%	32/32
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	3.99713 ± 0.2393	1319832.5000 ± 79436.8863	20.0%	0.0119 ± 0.0084	4102.4000 ± 2792.8856	0.2828 ± 0.0127	95600.0000 ± 3039.7368	100.0%	32/32
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	3.59176 ± 0.6811	1190225.3000 ± 218482.7231	70.0%	0.0116 ± 0.0053	3824.0000 ± 1580.0028	0.2137 ± 0.0136	72500.0000 ± 3170.1735	100.0%	32/34
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	3.52656 ± 0.8051	1176102.4000 ± 263042.2913	70.0%	0.0154 ± 0.0063	5068.8000 ± 1926.3449	0.1860 ± 0.0125	64100.0000 ± 3590.2646	100.0%	32/32
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	1.76258 ± 0.6522	575884.7000 ± 212264.7682	100.0%	0.0102 ± 0.0035	3479.4000 ± 1287.8182	0.1818 ± 0.0161	60500.0000 ± 2291.2878	100.0%	32/32
Q-learning ($\rho = 0.5, \epsilon = 1$)	1.20305 ± 0.1918	405112.7000 ± 60689.9822	100.0%	0.0051 ± 0.0044	1613.7000 ± 1544.9670	0.6304 ± 0.0550	213900.0000 ± 19684.7657	100.0%	32/32
Q-learning ($\rho = 0.7, \epsilon = 1$)	1.02061 ± 0.1794	344552.6000 ± 62568.4433	100.0%	0.0085 ± 0.0076	2666.1000 ± 2584.2665	0.4529 ± 0.0719	155300.0000 ± 21909.1305	100.0%	32/32
Q-learning ($\rho = 0.9, \epsilon = 1$)	0.96460 ± 0.2231	324083.5000 ± 71469.0675	100.0%	0.0064 ± 0.0065	2042.1000 ± 1891.6086	0.4776 ± 0.0802	160000.0000 ± 26164.8619	100.0%	32/32
Q-learning ($\rho = 0.99, \epsilon = 1$)	0.96659 ± 0.2904	319446.0000 ± 96412.0417	100.0%	0.0110 ± 0.0095	3823.2000 ± 3389.0056	0.3749 ± 0.0847	123500.0000 ± 24422.3259	100.0%	32/32
Async Value Iteration	0.09035 ± 0.0026	125114.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	32/32
Value Iteration	0.11383 ± 0.0066	154396.0000 ± 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 ρ 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 ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.37247 ± 0.0122	114617.4000 ± 187.8751	0.0%	0.0053 ± 0.0009	1517.0000 ± 162.5534	0.1398 ± 0.0067	43000.0000 ± 2280.3509	100.0%	31/31
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.35428 ± 0.0076	109239.9000 ± 623.2713	0.0%	0.0045 ± 0.0005	1506.9000 ± 227.1913	0.0986 ± 0.0083	30700.0000 ± 2100.0000	100.0%	31/33
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.37097 ± 0.0538	107828.4000 ± 1384.4830	0.0%	0.0050 ± 0.0007	1494.9000 ± 151.1717	0.1082 ± 0.0769	33200.0000 ± 23990.8316	100.0%	31/100002
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.35672 ± 0.0116	108514.1000 ± 1709.4866	0.0%	0.0056 ± 0.0012	1559.6000 ± 204.3880	0.1860 ± 0.0609	56600.0000 ± 18704.0103	100.0%	31/296
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.46290 ± 0.0091	147204.8000 ± 297.3829	0.0%	0.0061 ± 0.0016	1893.2000 ± 277.2716	0.2288 ± 0.0776	73857.1429 ± 25881.2043	70.0%	31/31
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.46012 ± 0.0137	146311.8000 ± 289.5506	0.0%	0.0055 ± 0.0011	1711.8000 ± 268.4846	0.2132 ± 0.0113	66000.0000 ± 29588.8493	40.0%	31/34
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	0.43303 ± 0.0106	138488.2000 ± 285.2051	0.0%	0.0060 ± 0.0011	1913.9000 ± 250.9874	0.3568 ± 0.0631	114500.0000 ± 16500.0000	20.0%	31/33
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	0.44265 ± 0.0125	137993.9000 ± 306.2262	0.0%	0.0052 ± 0.0009	1595.2000 ± 171.7066	nan ± nan	nan ± nan	0.0%	31/34
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.66724 ± 0.0156	214887.0000 ± 1162.4659	0.0%	0.0076 ± 0.0018	2378.4000 ± 657.2575	0.3220 ± 0.1674	105600.0000 ± 56873.8956	50.0%	31/33
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.65335 ± 0.0154	209485.3000 ± 618.2669	0.0%	0.0066 ± 0.0009	2187.1000 ± 314.9376	0.3104 ± 0.0783	102250.0000 ± 26309.4565	40.0%	31/32
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	0.64415 ± 0.0115	207142.8000 ± 821.3853	0.0%	0.0075 ± 0.0024	2378.1000 ± 482.4167	nan ± nan	nan ± nan	0.0%	31/33
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.65736 ± 0.0098	208470.3000 ± 881.3810	0.0%	0.0070 ± 0.0022	2325.8000 ± 714.1581	nan ± nan	nan ± nan	0.0%	31/33
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	1.29698 ± 0.0385	430024.7000 ± 2972.1627	0.0%	0.0081 ± 0.0027	2714.4000 ± 799.0636	0.6333 ± 0.2917	211000.0000 ± 97382.8967	70.0%	31/32
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	1.30140 ± 0.0282	427028.4000 ± 2753.3805	0.0%	0.0085 ± 0.0029	2850.4000 ± 849.9456	0.9404 ± 0.1961	311400.0000 ± 61856.6084	50.0%	31/33
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	1.30231 ± 0.0223	429349.7000 ± 2556.0207	0.0%	0.0078 ± 0.0016	2563.7000 ± 461.9909	nan ± nan	nan ± nan	0.0%	31/33
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	1.33780 ± 0.0157	444145.0000 ± 1647.8951	0.0%	0.0102 ± 0.0026	3366.3000 ± 981.7009	nan ± nan	nan ± nan	0.0%	31/32
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	3.22848 ± 0.0418	1095264.3000 ± 11150.3553	0.0%	0.0094 ± 0.0059	3136.1000 ± 1884.6837	1.2835 ± 0.6059	438500.0000 ± 204011.6418	80.0%	31/33
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	3.27396 ± 0.0773	1101426.5000 ± 10878.4900	0.0%	0.0079 ± 0.0044	2325.1000 ± 1158.6940	1.5485 ± 0.9478	518000.0000 ± 317805.4961	60.0%	31/32
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	3.45066 ± 0.0453	1168286.4000 ± 10096.5994	0.0%	0.0083 ± 0.0050	3044.8000 ± 1847.5530	2.5964 ± 0.0130	885000.0000 ± 17000.0000	20.0%	31/32
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	3.70953 ± 0.0484	1253590.4000 ± 14448.1124	0.0%	0.0090 ± 0.0026	2996.3000 ± 920.3565	nan ± nan	nan ± nan	0.0%	31/34
Q-learning ($\rho = 0.5, \epsilon = 1$)	13.94533 ± 0.1766	4764212.8000 ± 54616.4952	0.0%	0.0050 ± 0.0042	1618.4000 ± 1324.5059	3.0011 ± 3.1367	1022900.0000 ± 1070787.3225	100.0%	31/32
Q-learning ($\rho = 0.7, \epsilon = 1$)	13.99891 ± 0.2207	4795089.2000 ± 51372.6601	0.0%	0.0036 ± 0.0021	1223.7000 ± 618.9092	2.7212 ± 1.9902	932100.0000 ± 680721.5951	100.0%	31/33
Q-learning ($\rho = 0.9, \epsilon = 1$)	13.84819 ± 0.1945	4760311.1000 ± 47803.9988	0.0%	0.0067 ± 0.0065	2251.4000 ± 2306.8593	4.5475 ± 3.5510	1559333.3333 ± 1216161.6760	60.0%	31/32
Q-learning ($\rho = 0.99, \epsilon = 1$)	13.91682 ± 0.1248	4774551.0000 ± 32997.9805	0.0%	0.0060 ± 0.0057	2078.2000 ± 1974.2113	nan ± nan	nan ± nan	0.0%	31/33
Async Value Iteration	0.10508 ± 0.0038	146190.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	31/32
Value Iteration	0.13854 ± 0.0051	188718.0000 ± 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 ρ 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 ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.36513 ± 0.0125	112859.7000 ± 116.2050	0.0%	0.0042 ± 0.0005	1356.7000 ± 84.9883	0.1281 ± 0.0053	39500.0000 ± 500.0000	100.0%	31/4396
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.34452 ± 0.0094	106654.6000 ± 102.1237	0.0%	0.0043 ± 0.0004	1315.1000 ± 9.7514	0.0904 ± 0.0045	28000.0000 ± 0.0000	100.0%	31/31
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.33664 ± 0.0124	103423.9000 ± 88.7699	0.0%	0.0047 ± 0.0009	1319.7000 ± 31.8906	0.0716 ± 0.0053	21800.0000 ± 400.0000	100.0%	31/31
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.33875 ± 0.0142	102853.2000 ± 145.6130	0.0%	0.0045 ± 0.0005	1359.6000 ± 82.8869	0.0618 ± 0.0022	19000.0000 ± 0.0000	100.0%	31/33
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.46203 ± 0.0119	144659.5000 ± 361.3915	0.0%	0.0063 ± 0.0017	1902.5000 ± 412.3419	0.1402 ± 0.0063	44000.0000 ± 0.0000	100.0%	31/31
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.44829 ± 0.0182	139438.5000 ± 349.4439	0.0%	0.0057 ± 0.0008	1768.3000 ± 253.4269	0.1018 ± 0.0061	31800.0000 ± 400.0000	100.0%	31/31
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	0.43783 ± 0.0168	137089.0000 ± 610.6595	0.0%	0.0057 ± 0.0012	1725.5000 ± 282.0455	0.0796 ± 0.0072	25000.0000 ± 0.0000	100.0%	31/31
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	0.44168 ± 0.0112	138299.0000 ± 403.0796	0.0%	0.0058 ± 0.0011	1687.1000 ± 277.1907	0.0711 ± 0.0042	22200.0000 ± 400.0000	100.0%	31/31
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.65731 ± 0.0154	212269.9000 ± 1550.9998	0.0%	0.0061 ± 0.0006	1903.4000 ± 192.7123	0.1594 ± 0.0072	51000.0000 ± 632.4555	100.0%	31/31
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.65121 ± 0.0184	210603.8000 ± 1186.9124	0.0%	0.0062 ± 0.0016	2094.3000 ± 507.0217	0.1167 ± 0.0039	37900.0000 ± 300.0000	100.0%	31/31
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	0.65888 ± 0.0187	211653.4000 ± 810.7327	0.0%	0.0068 ± 0.0019	2121.1000 ± 363.6071	0.0955 ± 0.0037	31100.0000 ± 700.0000	100.0%	31/31
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.66664 ± 0.0147	215614.9000 ± 1171.6542	0.0%	0.0075 ± 0.0018	2346.2000 ± 651.0593	0.0845 ± 0.0035	27700.0000 ± 458.2576	100.0%	31/31
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	1.32853 ± 0.0290	437144.7000 ± 4931.2188	0.0%	0.0083 ± 0.0032	2612.4000 ± 728.0974	0.2014 ± 0.0109	66000.0000 ± 1612.4515	100.0%	31/32
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	1.37867 ± 0.0164	456752.4000 ± 3404.0699	0.0%	0.0099 ± 0.0037	3033.4000 ± 716.6369	0.1510 ± 0.0056	50200.0000 ± 979.7959	100.0%	31/31
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	1.43790 ± 0.0255	470740.2000 ± 3139.9116	0.0%	0.0087 ± 0.0027	2831.7000 ± 830.6499	0.1275 ± 0.0072	42100.0000 ± 1920.9373	100.0%	31/32
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	1.47902 ± 0.0222	493238.9000 ± 2277.8657	0.0%	0.0084 ± 0.0018	2776.1000 ± 678.2903	0.1161 ± 0.0044	38400.0000 ± 1356.4660	100.0%	31/31
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	2.28952 ± 0.5074	775214.8000 ± 171212.5264	100.0%	0.0096 ± 0.0054	3032.0000 ± 1662.5293	0.2548 ± 0.0092	86900.0000 ± 2981.6103	100.0%	31/31
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	1.93763 ± 0.6133	651227.4000 ± 201436.2162	100.0%	0.0067 ± 0.0044	2413.5000 ± 1718.9784	0.2040 ± 0.0151	68600.0000 ± 4127.9535	100.0%	31/32
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	1.41624 ± 0.4372	472450.5000 ± 142691.4930	100.0%	0.0107 ± 0.0068	3728.7000 ± 2238.1539	0.1762 ± 0.0087	58700.0000 ± 2100.0000	100.0%	31/31
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	0.93309 ± 0.2651	316115.5000 ± 90645.6125	100.0%	0.0076 ± 0.0030	2368.9000 ± 1017.2718	0.1617 ± 0.0099	54200.0000 ± 4261.4552	100.0%	31/33
Q-learning ($\rho = 0.5, \epsilon = 1$)	0.74431 ± 0.1423	258667.4000 ± 46697.7274	100.0%	0.0046 ± 0.0023	1498.6000 ± 789.8309	0.5014 ± 0.0546	174800.0000 ± 17503.1426	100.0%	31/32
Q-learning ($\rho = 0.7, \epsilon = 1$)	0.59378 ± 0.1131	205227.6000 ± 35179.7023	100.0%	0.0044 ± 0.0021	1289.9000 ± 793.1236	0.3684 ± 0.0589	127200.0000 ± 16587.9474	100.0%	31/31
Q-learning ($\rho = 0.9, \epsilon = 1$)	0.57472 ± 0.1143	199051.8000 ± 36223.4234	100.0%	0.0030 ± 0.0019	1086.3000 ± 796.4359	0.3153 ± 0.0466	110300.0000 ± 15691.0803	100.0%	31/33
Q-learning ($\rho = 0.99, \epsilon = 1$)	0.46894 ± 0.0688	160408.3000 ± 22891.0348	100.0%	0.0045 ± 0.0032	1582.0000 ± 1197.7031	0.2889 ± 0.0562	99200.0000 ± 18454.2678	100.0%	31/31
Async Value Iteration	0.08566 ± 0.0025	122268.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	31/31
Value Iteration	0.11020 ± 0.0053	151506.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	31/31

Table 41. Comparing performance of Q-learning as we change the learning rate ρ 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 ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.33427 ± 0.0122	105125.7000 ± 168.1048	0.0%	0.0033 ± 0.0005	1107.3000 ± 147.2610	0.0960 ± 0.0077	30100.0000 ± 1640.1219	100.0%	29/31
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.32748 ± 0.0138	101730.0000 ± 1430.0034	0.0%	0.0030 ± 0.0006	1015.1000 ± 129.7131	0.0700 ± 0.0045	21700.0000 ± 1100.0000	100.0%	29/31
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.32981 ± 0.0112	102508.9000 ± 1459.2476	0.0%	0.0032 ± 0.0007	1013.0000 ± 145.9377	0.0863 ± 0.0586	27600.0000 ± 18900.7936	100.0%	29/37
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.33552 ± 0.0117	104173.0000 ± 1482.8527	0.0%	0.0034 ± 0.0012	1029.0000 ± 118.8242	0.1318 ± 0.0789	40888.8889 ± 25431.5834	90.0%	29/35
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.43405 ± 0.0090	138928.2000 ± 499.2035	0.0%	0.0040 ± 0.0011	1429.5000 ± 311.4085	0.1782 ± 0.0679	57333.3333 ± 21664.1024	90.0%	29/31
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.41609 ± 0.0110	133855.8000 ± 361.0805	0.0%	0.0045 ± 0.0008	1314.9000 ± 212.1652	0.2415 ± 0.1086	77428.5714 ± 35362.8423	70.0%	29/30
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	0.40879 ± 0.0130	131461.9000 ± 483.5714	0.0%	0.0036 ± 0.0011	1102.2000 ± 309.2872	nan ± nan	nan ± nan	0.0%	29/31
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	0.40781 ± 0.0077	131072.9000 ± 596.3297	0.0%	0.0042 ± 0.0008	1219.2000 ± 187.8818	nan ± nan	nan ± nan	0.0%	29/31
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.64501 ± 0.0140	210120.2000 ± 559.0620	0.0%	0.0046 ± 0.0013	1592.5000 ± 523.0178	0.5552 ± 0.0808	181600.0000 ± 28302.6501	50.0%	29/31
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.62716 ± 0.0089	204995.1000 ± 479.7298	0.0%	0.0048 ± 0.0013	1485.7000 ± 397.7974	0.3024 ± 0.1961	100000.0000 ± 66000.0000	20.0%	29/33
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	0.62131 ± 0.0105	202853.4000 ± 1130.9142	0.0%	0.0049 ± 0.0017	1574.0000 ± 500.7021	nan ± nan	nan ± nan	0.0%	29/31
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.62282 ± 0.0129	203244.8000 ± 818.0984	0.0%	0.0044 ± 0.0012	1410.1000 ± 314.7696	nan ± nan	nan ± nan	0.0%	29/31
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	1.35667 ± 0.0148	459026.0000 ± 2101.6706	0.0%	0.0069 ± 0.0037	2345.3000 ± 1079.9207	0.5004 ± 0.2531	169428.5714 ± 85780.2127	70.0%	29/29
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	1.35292 ± 0.0245	459012.2000 ± 1877.8213	0.0%	0.0066 ± 0.0034	2197.4000 ± 1178.8474	0.6868 ± 0.2734	229750.0000 ± 94221.4811	40.0%	29/30
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	1.34463 ± 0.0268	455792.3000 ± 1940.4911	0.0%	0.0078 ± 0.0037	2449.2000 ± 1051.0859	0.7965 ± 0.0000	266000.0000 ± 0.0000	10.0%	29/31
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	1.35708 ± 0.0132	459658.1000 ± 1821.0703	0.0%	0.0073 ± 0.0040	2472.1000 ± 1381.1790	nan ± nan	nan ± nan	0.0%	29/29
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	4.10222 ± 0.0447	1416290.2000 ± 13406.2110	0.0%	0.0072 ± 0.0058	2373.9000 ± 1710.7167	0.6198 ± 0.5014	214400.0000 ± 173779.8607	100.0%	29/30
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	4.09371 ± 0.0562	1403508.6000 ± 16330.4633	0.0%	0.0075 ± 0.0041	2689.1000 ± 1621.3430	1.0343 ± 0.5457	356100.0000 ± 182611.3085	100.0%	29/31
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	4.16473 ± 0.0605	1425769.6000 ± 9162.1709	0.0%	0.0082 ± 0.0045	2627.8000 ± 1332.1414	1.6117 ± 1.1108	550250.0000 ± 378253.8004	80.0%	29/30
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	4.22978 ± 0.0746	1441110.6000 ± 14152.2690	0.0%	0.0089 ± 0.0055	2772.5000 ± 1824.4263	nan ± nan	nan ± nan	0.0%	29/30
Q-learning ($\rho = 0.5, \epsilon = 1$)	17.30241 ± 0.1672	5971427.6000 ± 66264.1556	0.0%	0.0105 ± 0.0099	3899.2000 ± 4044.8257	0.8476 ± 0.3117	296100.0000 ± 105706.6233	100.0%	29/30
Q-learning ($\rho = 0.7, \epsilon = 1$)	17.21702 ± 0.2234	5933093.4000 ± 54805.6557	0.0%	0.0050 ± 0.0045	1572.8000 ± 1801.3915	0.9488 ± 0.6939	325300.0000 ± 240383.4645	100.0%	29/29
Q-learning ($\rho = 0.9, \epsilon = 1$)	17.40180 ± 0.1732	5991735.3000 ± 40936.1980	0.0%	0.0071 ± 0.0037	2606.7000 ± 1454.0691	2.0168 ± 1.9017	698400.0000 ± 659821.2182	100.0%	29/33
Q-learning ($\rho = 0.99, \epsilon = 1$)	17.29445 ± 0.1350	5979980.7000 ± 45321.9834	0.0%	0.0079 ± 0.0051	2628.2000 ± 1748.4152	14.1482 ± 0.0000	4837000.0000 ± 0.0000	10.0%	29/29
Async Value Iteration	0.10354 ± 0.0032	153272.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	29/32
Value Iteration	0.15642 ± 0.0028	228344.0000 ± 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 ρ 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 ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.33216 ± 0.0107	103263.5000 ± 97.8951	0.0%	0.0034 ± 0.0004	991.9000 ± 53.0838	0.0905 ± 0.0071	27900.0000 ± 300.0000	100.0%	29/29
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.31104 ± 0.0058	98369.3000 ± 169.5140	0.0%	0.0038 ± 0.0006	1050.7000 ± 4.9000	0.0637 ± 0.0026	20100.0000 ± 300.0000	100.0%	29/29
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.30551 ± 0.0091	96413.8000 ± 1629.3815	0.0%	0.0033 ± 0.0005	1046.6000 ± 41.6370	0.0503 ± 0.0020	16000.0000 ± 0.0000	100.0%	29/31
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.30452 ± 0.0095	95497.2000 ± 209.7593	0.0%	0.0032 ± 0.0007	1015.1000 ± 83.5146	0.0477 ± 0.0031	14700.0000 ± 458.2576	100.0%	29/30
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.42343 ± 0.0088	136151.9000 ± 548.4530	0.0%	0.0047 ± 0.0008	1411.1000 ± 257.0659	0.0966 ± 0.0037	30800.0000 ± 400.0000	100.0%	29/31
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.40466 ± 0.0088	130875.1000 ± 687.4162	0.0%	0.0046 ± 0.0040	1373.0000 ± 397.5077	0.0711 ± 0.0023	22600.0000 ± 489.8979	100.0%	29/29
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	0.40134 ± 0.0112	128105.8000 ± 324.8128	0.0%	0.0039 ± 0.0007	1251.4000 ± 130.5214	0.0579 ± 0.0032	18100.0000 ± 300.0000	100.0%	29/29
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	0.39706 ± 0.0101	127570.9000 ± 316.5042	0.0%	0.0039 ± 0.0009	1311.8000 ± 203.7694	0.0518 ± 0.0015	16800.0000 ± 400.0000	100.0%	29/29
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.63257 ± 0.0182	205580.9000 ± 1404.5667	0.0%	0.0052 ± 0.0019	1622.2000 ± 578.2848	0.1113 ± 0.0068	35800.0000 ± 600.0000	100.0%	29/29
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.62017 ± 0.0214	200313.1000 ± 1089.1233	0.0%	0.0056 ± 0.0015	1697.9000 ± 377.1388	0.0845 ± 0.0060	27200.0000 ± 600.0000	100.0%	29/30
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	0.60459 ± 0.0116	199227.1000 ± 805.0397	0.0%	0.0041 ± 0.0013	1376.9000 ± 330.9065	0.0686 ± 0.0038	22500.0000 ± 500.0000	100.0%	29/29
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.60542 ± 0.0120	199838.4000 ± 1353.3196	0.0%	0.0054 ± 0.0017	1721.2000 ± 528.9037	0.0623 ± 0.0038	20200.0000 ± 600.0000	100.0%	29/29
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	1.35722 ± 0.0182	455451.6000 ± 3052.3085	0.0%	0.0063 ± 0.0040	2006.3000 ± 1203.4328	0.1411 ± 0.0052	47300.0000 ± 900.0000	100.0%	29/29
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	1.33913 ± 0.0179	451691.6000 ± 3530.7195	0.0%	0.0060 ± 0.0035	2007.4000 ± 1127.2885	0.1116 ± 0.0055	37000.0000 ± 1095.4451	100.0%	29/29
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	1.37123 ± 0.0189	460888.6000 ± 2646.9526	0.0%	0.0073 ± 0.0028	2533.5000 ± 1178.8222	0.0933 ± 0.0061	31900.0000 ± 1135.7817	100.0%	29/30
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	1.37888 ± 0.0150	467527.8000 ± 3335.9965	0.0%	0.0052 ± 0.0020	1786.7000 ± 533.3095	0.0851 ± 0.0035	28400.0000 ± 663.3250	100.0%	29/30
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	4.15838 ± 0.0608	1426565.1000 ± 18949.2330	10.0%	0.0072 ± 0.0041	2383.6000 ± 1194.7452	0.2110 ± 0.0190	70900.0000 ± 5374.9419	100.0%	29/29
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	3.40310 ± 0.7200	1168608.8000 ± 239906.4924	80.0%	0.0053 ± 0.0045	1952.8000 ± 1346.8030	0.1548 ± 0.0105	52900.0000 ± 3986.2263	100.0%	29/29
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	2.64154 ± 0.4819	907039.3000 ± 167557.7797	100.0%	0.0071 ± 0.0051	2274.9000 ± 1836.3152	0.1330 ± 0.0121	45500.0000 ± 2692.5824	100.0%	29/29
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	1.76411 ± 0.4392	606533.8000 ± 153380.7036	100.0%	0.0048 ± 0.0020	1577.6000 ± 557.6015	0.1184 ± 0.0113	41700.0000 ± 2609.5977	100.0%	29/30
Q-learning ($\rho = 0.5, \epsilon = 1$)	3.02736 ± 0.3939	1044897.5000 ± 133304.0531	100.0%	0.0050 ± 0.0037	1883.4000 ± 1331.9621	0.6006 ± 0.0763	207400.0000 ± 24670.6303	100.0%	29/31
Q-learning ($\rho = 0.7, \epsilon = 1$)	2.60679 ± 0.5753	901460.4000 ± 202265.4853	100.0%	0.0082 ± 0.0077	2664.7000 ± 2314.6502	0.4190 ± 0.0814	144900.0000 ± 27005.3698	100.0%	29/29
Q-learning ($\rho = 0.9, \epsilon = 1$)	1.98026 ± 0.3646	680915.2000 ± 125997.0656	100.0%	0.0108 ± 0.0084	3578.7000 ± 2836.1716	0.3622 ± 0.0473	122700.0000 ± 14036.0251	100.0%	29/29
Q-learning ($\rho = 0.99, \epsilon = 1$)	2.26373 ± 0.4705	757744.1000 ± 151013.5598	100.0%	0.0101 ± 0.0096	3579.8000 ± 3820.0271	0.3187 ± 0.0480	109000.0000 ± 16619.2659	100.0%	29/29
Async Value Iteration	0.08177 ± 0.0028	125120.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	29/29
Value Iteration	0.12583 ± 0.0071	181424.0000 ± 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 ρ 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 ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.04862 ± 0.0022	15439.8000 ± 22.0082	0.0%	0.0000 ± 0.0000	44.1000 ± 5.5036	0.0040 ± 0.0014	1200.0000 ± 400.0000	100.0%	6/6
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.04577 ± 0.0038	14974.1000 ± 1325.0927	10.0%	0.0000 ± nan	39.4000 ± 1.6852	0.0034 ± 0.0009	1100.0000 ± 300.0000	100.0%	6/6
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.05169 ± 0.0036	16332.6000 ± 1134.5385	0.0%	0.0010 ± 0.0000	38.4000 ± 1.8000	0.0035 ± 0.0008	1000.0000 ± 0.0000	100.0%	6/8
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.01597 ± 0.0203	5147.4000 ± 6494.3600	80.0%	0.0012 ± 0.0000	41.0000 ± 2.8284	0.0034 ± 0.0011	1100.0000 ± 300.0000	100.0%	6/8
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.01654 ± 0.0050	5202.5000 ± 1536.4906	100.0%	0.0010 ± 0.0000	48.3000 ± 9.7370	0.0041 ± 0.0015	1300.0000 ± 458.2576	100.0%	6/6
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.01109 ± 0.0060	3603.8000 ± 2059.5251	100.0%	0.0010 ± 0.0000	47.1000 ± 11.9202	0.0036 ± 0.0006	1000.0000 ± 0.0000	100.0%	6/6
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	0.00631 ± 0.0031	2002.8000 ± 895.1018	100.0%	0.0010 ± 0.0000	49.1000 ± 8.3241	0.0031 ± 0.0007	1000.0000 ± 0.0000	100.0%	6/7
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	0.00462 ± 0.0020	1404.8000 ± 662.3947	100.0%	0.0010 ± 0.0000	47.5000 ± 3.6946	0.0039 ± 0.0012	1100.0000 ± 300.0000	100.0%	6/6
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.00857 ± 0.0029	3004.4000 ± 1001.0057	100.0%	0.0010 ± 0.0000	53.9000 ± 18.9813	0.0034 ± 0.0012	1200.0000 ± 400.0000	100.0%	6/6
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.00556 ± 0.0019	1803.7000 ± 600.2493	100.0%	0.0010 ± 0.0000	64.7000 ± 12.9310	0.0032 ± 0.0009	1100.0000 ± 300.0000	100.0%	6/7
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	0.00613 ± 0.0032	2107.8000 ± 1220.1192	100.0%	0.0010 ± 0.0000	47.2000 ± 8.5065	0.0035 ± 0.0011	1200.0000 ± 400.0000	100.0%	6/6
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.00371 ± 0.0014	1304.1000 ± 457.7640	100.0%	0.0010 ± 0.0000	55.8000 ± 15.1578	0.0030 ± 0.0000	1000.0000 ± 0.0000	100.0%	6/6
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	0.00825 ± 0.0030	2613.4000 ± 800.3729	100.0%	0.0000 ± 0.0000	62.2000 ± 12.9445	0.0036 ± 0.0015	1100.0000 ± 300.0000	100.0%	6/8
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	0.00610 ± 0.0023	2014.0000 ± 638.2051	100.0%	0.0013 ± 0.0004	58.4000 ± 22.2225	0.0033 ± 0.0012	1100.0000 ± 300.0000	100.0%	6/7
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	0.00430 ± 0.0019	1310.4000 ± 461.7654	100.0%	0.0010 ± 0.0000	42.9000 ± 27.5770	0.0031 ± 0.0010	1000.0000 ± 0.0000	100.0%	6/8
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	0.00348 ± 0.0012	1211.8000 ± 404.1685	100.0%	0.0010 ± 0.0000	48.4000 ± 22.7473	0.0030 ± 0.0005	1000.0000 ± 0.0000	100.0%	6/6
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	0.00812 ± 0.0030	3022.0000 ± 1180.4918	100.0%	0.0010 ± 0.0000	29.3000 ± 12.1083	0.0032 ± 0.0012	1100.0000 ± 300.0000	100.0%	6/6
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	0.00472 ± 0.0034	1617.3000 ± 796.8305	100.0%	0.0010 ± 0.0000	56.1000 ± 15.6042	0.0031 ± 0.0007	1000.0000 ± 0.0000	100.0%	6/6
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	0.00632 ± 0.0032	2119.6000 ± 1143.0843	100.0%	0.0010 ± 0.0000	37.1000 ± 25.3316	0.0033 ± 0.0010	1100.0000 ± 300.0000	100.0%	6/6
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	0.00412 ± 0.0034	1420.0000 ± 1217.0481	100.0%	0.0000 ± 0.0000	44.6000 ± 30.2000	0.0030 ± 0.0004	1000.0000 ± 0.0000	100.0%	6/7
Q-learning ($\rho = 0.5, \epsilon = 1$)	0.01071 ± 0.0023	3549.3000 ± 790.8906	100.0%	0.0010 ± 0.0000	48.4000 ± 27.1926	0.0030 ± 0.0003	1000.0000 ± 0.0000	100.0%	6/6
Q-learning ($\rho = 0.7, \epsilon = 1$)	0.00690 ± 0.0020	2438.4000 ± 775.0199	100.0%	0.0009 ± 0.0000	46.8000 ± 33.4987	0.0032 ± 0.0006	1000.0000 ± 0.0000	100.0%	6/8
Q-learning ($\rho = 0.9, \epsilon = 1$)	0.00725 ± 0.0030	2455.4000 ± 1011.8974	100.0%	0.0010 ± 0.0000	76.6000 ± 45.0981	0.0034 ± 0.0009	1100.0000 ± 300.0000	100.0%	6/6
Q-learning ($\rho = 0.99, \epsilon = 1$)	0.00497 ± 0.0023	1736.9000 ± 760.3822	100.0%	0.0000 ± 0.0000	46.8000 ± 33.7485	0.0034 ± 0.0011	1100.0000 ± 300.0000	100.0%	6/6
Async Value Iteration	0.00405 ± 0.0001	1700.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	6/6
Value Iteration	0.00645 ± 0.0009	2516.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	6/6

Table 44. Comparing performance of Q-learning as we change the learning rate ρ 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 ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.04766 ± 0.0023	15213.5000 ± 10.1415	0.0%	0.0010 ± 0.0000	39.4000 ± 1.2000	0.0032 ± 0.0004	1000.0000 ± 0.0000	100.0%	6/6
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.04846 ± 0.0017	15174.0000 ± 4.1473	0.0%	0.0000 ± 0.0000	39.6000 ± 1.8000	0.0034 ± 0.0007	1000.0000 ± 0.0000	100.0%	6/6
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.05209 ± 0.0035	15185.0000 ± 78.5519	0.0%	0.0010 ± 0.0000	39.0000 ± 0.0000	0.0035 ± 0.0005	1000.0000 ± 0.0000	100.0%	6/6
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.01063 ± 0.0052	3200.9000 ± 1536.6980	100.0%	0.0010 ± 0.0000	39.5000 ± 1.0247	0.0039 ± 0.0015	1200.0000 ± 600.0000	100.0%	6/7
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.01759 ± 0.0087	5703.0000 ± 2865.5200	100.0%	0.0010 ± 0.0000	50.5000 ± 9.7082	0.0033 ± 0.0007	1000.0000 ± 0.0000	100.0%	6/6
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.01216 ± 0.0064	4103.7000 ± 2342.9184	100.0%	0.0000 ± 0.0000	47.5000 ± 8.0654	0.0030 ± 0.0006	1000.0000 ± 0.0000	100.0%	6/6
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	0.00946 ± 0.0095	3002.1000 ± 3194.1199	100.0%	0.0012 ± 0.0002	47.4000 ± 8.2849	0.0035 ± 0.0006	1000.0000 ± 0.0000	100.0%	6/6
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	0.00529 ± 0.0035	1602.9000 ± 1019.7657	100.0%	0.0010 ± 0.0000	48.0000 ± 7.8102	0.0034 ± 0.0006	1000.0000 ± 0.0000	100.0%	6/6
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.00905 ± 0.0076	2905.5000 ± 2425.9665	100.0%	0.0010 ± 0.0000	52.8000 ± 20.3263	0.0035 ± 0.0012	1100.0000 ± 300.0000	100.0%	6/6
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.00372 ± 0.0010	1306.9000 ± 456.8458	100.0%	0.0010 ± 0.0000	54.4000 ± 15.8569	0.0034 ± 0.0007	1000.0000 ± 0.0000	100.0%	6/6
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	0.00515 ± 0.0054	1506.5000 ± 1206.1261	100.0%	0.0010 ± 0.0000	46.8000 ± 11.8474	0.0030 ± 0.0005	1000.0000 ± 0.0000	100.0%	6/6
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.00411 ± 0.0030	1406.5000 ± 915.7572	100.0%	0.0000 ± 0.0000	56.2000 ± 10.8517	0.0031 ± 0.0003	1000.0000 ± 0.0000	100.0%	6/6
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	0.00619 ± 0.0027	1912.7000 ± 702.8627	100.0%	0.0000 ± 0.0000	65.1000 ± 26.7000	0.0033 ± 0.0007	1000.0000 ± 0.0000	100.0%	6/6
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	0.00425 ± 0.0016	1410.6000 ± 487.8080	100.0%	0.0010 ± 0.0000	63.2000 ± 31.5144	0.0032 ± 0.0004	1000.0000 ± 0.0000	100.0%	6/6
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	0.00386 ± 0.0021	1218.3000 ± 596.5878	100.0%	0.0000 ± 0.0000	50.4000 ± 21.1858	0.0032 ± 0.0011	1000.0000 ± 0.0000	100.0%	6/6
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	0.00419 ± 0.0019	1309.7000 ± 455.9820	100.0%	0.0010 ± 0.0000	44.0000 ± 26.6871	0.0033 ± 0.0011	1100.0000 ± 300.0000	100.0%	6/6
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	0.00768 ± 0.0027	2624.7000 ± 918.9460	100.0%	0.0010 ± 0.0000	58.6000 ± 25.8116	0.0029 ± 0.0003	1000.0000 ± 0.0000	100.0%	6/7
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	0.00620 ± 0.0027	2027.3000 ± 884.4308	100.0%	0.0010 ± 0.0000	48.2000 ± 27.4401	0.0031 ± 0.0005	1000.0000 ± 0.0000	100.0%	6/6
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	0.00440 ± 0.0023	1416.8000 ± 662.1282	100.0%	0.0000 ± 0.0000	55.3000 ± 31.4517	0.0031 ± 0.0007	1000.0000 ± 0.0000	100.0%	6/8
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	0.00406 ± 0.0023	1336.2000 ± 471.4076	100.0%	0.0010 ± 0.0000	38.8000 ± 18.9832	0.0029 ± 0.0006	1000.0000 ± 0.0000	100.0%	6/6
Q-learning ($\rho = 0.5, \epsilon = 1$)	0.01071 ± 0.0033	3640.6000 ± 914.5519	100.0%	0.0010 ± 0.0000	24.0000 ± 12.2393	0.0033 ± 0.0017	1100.0000 ± 300.0000	100.0%	6/6
Q-learning ($\rho = 0.7, \epsilon = 1$)	0.00691 ± 0.0027	2357.1000 ± 906.2989	100.0%	0.0010 ± 0.0000	46.4000 ± 30.9587	0.0031 ± 0.0003	1000.0000 ± 0.0000	100.0%	6/6
Q-learning ($\rho = 0.9, \epsilon = 1$)	0.00631 ± 0.0026	2130.9000 ± 832.7856	100.0%	0.0010 ± 0.0000	43.8000 ± 42.2488	0.0033 ± 0.0010	1100.0000 ± 300.0000	100.0%	6/6
Q-learning ($\rho = 0.99, \epsilon = 1$)	0.00321 ± 0.0016	1123.8000 ± 297.6692	100.0%	0.0010 ± 0.0000	39.0000 ± 27.1735	0.0028 ± 0.0006	1000.0000 ± 0.0000	100.0%	6/6
Async Value Iteration	0.00325 ± 0.0004	1224.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	6/6
Value Iteration	0.00489 ± 0.0005	1836.0000 ± 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 ρ 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 ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.83249 ± 0.0195	264779.6000 ± 263.9455	0.0%	0.0048 ± 0.0013	1665.4000 ± 512.5989	0.2256 ± 0.0078	71100.0000 ± 2947.8806	100.0%	73/77
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.83544 ± 0.0599	255190.2000 ± 2224.9005	0.0%	0.0047 ± 0.0024	1359.1000 ± 475.6046	0.1622 ± 0.0120	49400.0000 ± 2615.3394	100.0%	73/100002
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.82218 ± 0.0200	261503.5000 ± 3559.2702	0.0%	0.0047 ± 0.0021	1303.8000 ± 598.0062	0.2267 ± 0.1634	72666.6667 ± 53441.5569	90.0%	75/210
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.84797 ± 0.0197	268361.8000 ± 2379.4391	0.0%	0.0040 ± 0.0012	1213.7000 ± 379.1628	0.2129 ± 0.0995	67000.0000 ± 32000.0000	20.0%	78/333
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	1.09374 ± 0.0134	354101.8000 ± 595.4437	0.0%	0.0061 ± 0.0028	1890.3000 ± 947.1300	0.3063 ± 0.0865	98900.0000 ± 26032.4797	100.0%	73/77
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	1.06019 ± 0.0180	343039.1000 ± 667.8831	0.0%	0.0048 ± 0.0041	1465.0000 ± 1181.4844	0.3598 ± 0.1795	117200.0000 ± 57875.3834	100.0%	73/77
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	1.05142 ± 0.0180	340876.0000 ± 678.3694	0.0%	0.0034 ± 0.0015	1193.4000 ± 514.9717	0.3378 ± 0.1401	109000.0000 ± 46566.0821	100.0%	73/2626
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	1.04891 ± 0.0169	343243.7000 ± 1504.2654	0.0%	0.0100 ± 0.0055	3310.6000 ± 1792.9804	0.3729 ± 0.2300	121200.0000 ± 75744.0427	100.0%	73/482
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	1.67280 ± 0.0291	542591.1000 ± 1717.5992	0.0%	0.0102 ± 0.0056	3367.6000 ± 1941.9648	0.5211 ± 0.1822	170300.0000 ± 59008.5587	100.0%	73/77
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	1.62994 ± 0.0320	532254.6000 ± 1110.4212	0.0%	0.0091 ± 0.0044	3012.5000 ± 1418.1769	0.3953 ± 0.1658	131400.0000 ± 52442.7307	100.0%	73/75
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	1.62375 ± 0.0113	532702.9000 ± 1277.0390	0.0%	0.0104 ± 0.0071	3473.3000 ± 2392.1237	0.5561 ± 0.2936	180500.0000 ± 93570.5616	100.0%	73/1215
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	1.70275 ± 0.1623	532676.4000 ± 1119.3726	0.0%	0.0081 ± 0.0037	2570.8000 ± 1323.9925	0.4862 ± 0.3775	147000.0000 ± 113822.6691	100.0%	73/1190
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	3.47534 ± 0.0379	1170012.2000 ± 5296.8401	0.0%	0.0202 ± 0.0082	6715.7000 ± 2639.9858	0.6041 ± 0.2729	203900.0000 ± 92684.8963	100.0%	73/76
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	3.46982 ± 0.0314	1169264.3000 ± 3635.3480	0.0%	0.0166 ± 0.0090	5640.8000 ± 2844.1041	0.4209 ± 0.1569	149900.0000 ± 54725.5882	100.0%	73/75
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	3.53218 ± 0.1073	1171866.5000 ± 2795.7716	0.0%	0.0154 ± 0.0091	4989.2000 ± 2665.4561	0.4960 ± 0.3073	166900.0000 ± 103700.0000	100.0%	73/134
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	3.52646 ± 0.0584	1174115.7000 ± 4024.4103	0.0%	0.0144 ± 0.0078	5075.6000 ± 2686.6880	0.9399 ± 0.8252	310200.0000 ± 262033.8909	100.0%	73/306
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	9.95711 ± 0.1040	3383447.3000 ± 13848.9996	0.0%	0.0373 ± 0.0210	12458.1000 ± 6668.1589	0.9189 ± 0.3297	314700.0000 ± 113260.8052	100.0%	73/75
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	10.06367 ± 0.0805	3405791.6000 ± 19143.3199	0.0%	0.0350 ± 0.0145	12192.7000 ± 4719.8426	0.8740 ± 0.3548	296700.0000 ± 122547.9906	100.0%	73/1184
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	10.12019 ± 0.1112	3429441.8000 ± 17535.3803	0.0%	0.0365 ± 0.0253	12366.9000 ± 8340.2219	0.9264 ± 0.3555	311400.0000 ± 118766.3252	100.0%	73/3031
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	10.07967 ± 0.0874	3425728.3000 ± 22238.1818	0.0%	0.0306 ± 0.0195	10482.2000 ± 6587.8251	0.8221 ± 0.4847	281200.0000 ± 167865.3031	100.0%	73/75
Q-learning ($\rho = 0.5, \epsilon = 1$)	25.73975 ± 0.1861	8858992.2000 ± 16766.6485	0.0%	0.1131 ± 0.1542	38799.6000 ± 53747.2304	2.2514 ± 0.6671	779100.0000 ± 227674.9657	100.0%	73/74
Q-learning ($\rho = 0.7, \epsilon = 1$)	25.85669 ± 0.2058	8866833.8000 ± 10524.3605	0.0%	0.1203 ± 0.1520	41029.9000 ± 50871.7990	2.0149 ± 0.5291	697000.0000 ± 181209.8231	100.0%	73/80
Q-learning ($\rho = 0.9, \epsilon = 1$)	25.93604 ± 0.2182	8866027.2000 ± 10242.8182	0.0%	0.1469 ± 0.0673	51328.8000 ± 24185.7968	1.8719 ± 0.6408	642100.0000 ± 216747.5259	100.0%	73/78
Q-learning ($\rho = 0.99, \epsilon = 1$)	28.05989 ± 1.7915	8866733.9000 ± 15794.0405	0.0%	0.1936 ± 0.1284	62289.1000 ± 42245.4083	1.4471 ± 0.2528	468700.0000 ± 95596.0773	100.0%	73/75
Async Value Iteration	0.08066 ± 0.0043	85158.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	73/77
Value Iteration	0.16040 ± 0.0039	180774.0000 ± 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 ρ 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 ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.82993 ± 0.0115	260739.3000 ± 48.6149	0.0%	0.0048 ± 0.0006	1444.3000 ± 106.8673	0.2228 ± 0.0073	68900.0000 ± 538.5165	100.0%	73/73
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.78513 ± 0.0115	248408.2000 ± 444.7241	0.0%	0.0046 ± 0.0005	1387.9000 ± 38.0301	0.1591 ± 0.0056	49300.0000 ± 458.2576	100.0%	73/75
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.78944 ± 0.0456	244120.9000 ± 1452.0210	0.0%	0.0038 ± 0.0010	1394.2000 ± 314.9079	0.1201 ± 0.0053	37800.0000 ± 400.0000	100.0%	73/17210
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.77546 ± 0.0195	243783.2000 ± 1594.2847	0.0%	0.0045 ± 0.0005	1388.3000 ± 25.8961	0.1078 ± 0.0034	33700.0000 ± 458.2576	100.0%	73/77
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	1.07721 ± 0.0084	348343.1000 ± 651.3309	0.0%	0.0073 ± 0.0034	2198.8000 ± 1011.6971	0.2361 ± 0.0089	75600.0000 ± 663.3250	100.0%	73/73
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	1.04090 ± 0.0155	337002.8000 ± 650.3260	0.0%	0.0048 ± 0.0040	1353.2000 ± 1037.7931	0.1665 ± 0.0042	55100.0000 ± 943.3981	100.0%	73/75
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	1.01295 ± 0.0125	331276.5000 ± 579.6087	0.0%	0.0035 ± 0.0011	1161.5000 ± 269.6792	0.1365 ± 0.0062	43700.0000 ± 640.3124	100.0%	73/74
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	1.01541 ± 0.0141	328914.1000 ± 564.6745	0.0%	0.0036 ± 0.0020	1099.6000 ± 609.2000	0.1227 ± 0.0087	39500.0000 ± 2061.5528	100.0%	73/73
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	1.62366 ± 0.0177	535649.3000 ± 1398.5836	0.0%	0.0077 ± 0.0038	2564.1000 ± 1306.3817	0.2717 ± 0.0077	86400.0000 ± 1113.5529	100.0%	73/75
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	1.59911 ± 0.0211	527009.5000 ± 1237.8385	0.0%	0.0094 ± 0.0062	2931.4000 ± 1956.7226	0.2041 ± 0.0055	65800.0000 ± 871.7798	100.0%	73/73
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	1.52830 ± 0.1005	502568.3000 ± 29572.5745	40.0%	0.0062 ± 0.0038	1965.0000 ± 1074.2367	0.1654 ± 0.0064	53600.0000 ± 800.0000	100.0%	73/73
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	1.42686 ± 0.1613	470006.9000 ± 52303.9108	60.0%	0.0065 ± 0.0031	2075.1000 ± 1091.0605	0.1506 ± 0.0060	49600.0000 ± 800.0000	100.0%	73/74
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	0.62042 ± 0.0825	208820.0000 ± 27455.5621	100.0%	0.0184 ± 0.0063	6124.3000 ± 2295.8321	0.3350 ± 0.0088	113500.0000 ± 1910.4973	100.0%	73/75
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	0.42993 ± 0.0514	144077.7000 ± 17788.1487	100.0%	0.0127 ± 0.0059	4394.3000 ± 2039.2830	0.2601 ± 0.0091	87200.0000 ± 1833.0303	100.0%	73/73
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	0.35981 ± 0.0913	120633.2000 ± 30433.5692	100.0%	0.0198 ± 0.0083	6789.5000 ± 2892.4318	0.2239 ± 0.0104	74900.0000 ± 2624.8809	100.0%	73/75
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	0.29206 ± 0.0452	97167.5000 ± 13056.1704	100.0%	0.0134 ± 0.0081	4290.8000 ± 2428.5683	0.2080 ± 0.0082	69200.0000 ± 2357.9652	100.0%	73/75
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	0.52948 ± 0.1420	181915.9000 ± 48434.5347	100.0%	0.0208 ± 0.0120	6859.9000 ± 4023.1146	0.4311 ± 0.0153	148400.0000 ± 5122.4994	100.0%	73/73
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	0.36716 ± 0.0216	124905.9000 ± 6861.0892	100.0%	0.0311 ± 0.0171	10556.1000 ± 5940.0527	0.3574 ± 0.0207	121500.0000 ± 5500.0000	100.0%	73/73
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	0.33836 ± 0.0547	118232.4000 ± 16603.1386	100.0%	0.0281 ± 0.0152	9606.0000 ± 4959.5078	0.3017 ± 0.0132	106100.0000 ± 4060.7881	100.0%	73/73
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	0.31871 ± 0.0554	110124.3000 ± 16029.4892	100.0%	0.0300 ± 0.0199	9923.9000 ± 6297.7764	0.3001 ± 0.0240	103900.0000 ± 6347.4404	100.0%	73/73
Q-learning ($\rho = 0.5, \epsilon = 1$)	2.99005 ± 0.6431	1037520.5000 ± 224749.2626	100.0%	0.1704 ± 0.1517	58870.9000 ± 50659.1635	2.3729 ± 0.3603	824700.0000 ± 129351.4979	100.0%	73/75
Q-learning ($\rho = 0.7, \epsilon = 1$)	3.03786 ± 1.2796	1046223.6000 ± 440196.1127	100.0%	0.0746 ± 0.0425	25056.4000 ± 13471.0723	1.9613 ± 0.3731	673400.0000 ± 122196.7266	100.0%	73/73
Q-learning ($\rho = 0.9, \epsilon = 1$)	2.05488 ± 0.8177	711559.7000 ± 284247.7162	100.0%	0.1607 ± 0.1877	55020.2000 ± 63214.5579	1.6575 ± 0.3488	574500.0000 ± 120845.5626	100.0%	73/75
Q-learning ($\rho = 0.99, \epsilon = 1$)	1.59392 ± 0.3695	550156.6000 ± 124750.5125	100.0%	0.0639 ± 0.0364	22644.3000 ± 13536.2447	1.4103 ± 0.3308	487300.0000 ± 112504.2666	100.0%	73/74
Async Value Iteration	0.05407 ± 0.0031	68724.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	73/73
Value Iteration	0.12356 ± 0.0066	152388.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	73/75

Table 47. Comparing performance of Q-learning as we change the learning rate ρ 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 ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	1.35049 ± 0.1266	349487.5000 ± 371.1585	0.0%	0.0154 ± 0.0096	3049.8000 ± 975.2360	0.4556 ± 0.0755	116300.0000 ± 2282.5424	100.0%	91/420
Q-learning ($\rho = 0.7, \epsilon = 0$)	1.21918 ± 0.0737	330855.7000 ± 1200.4179	0.0%	0.0093 ± 0.0027	2602.5000 ± 715.6272	0.3117 ± 0.0414	84100.0000 ± 3112.8765	100.0%	91/99
Q-learning ($\rho = 0.9, \epsilon = 0$)	1.18828 ± 0.0713	330840.3000 ± 2305.1930	0.0%	0.0092 ± 0.0034	2725.0000 ± 1024.8644	0.3035 ± 0.1566	86700.0000 ± 46456.5388	100.0%	93/99
Q-learning ($\rho = 0.99, \epsilon = 0$)	1.18778 ± 0.0537	334367.3000 ± 1915.6908	0.0%	0.0081 ± 0.0027	2486.8000 ± 815.5651	0.2972 ± 0.2505	85900.0000 ± 72693.1221	100.0%	96/100002
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	1.57223 ± 0.0228	452201.9000 ± 558.4995	0.0%	0.0116 ± 0.0089	3039.6000 ± 2594.0057	0.5420 ± 0.0761	154800.0000 ± 21198.1131	100.0%	91/614
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	1.49643 ± 0.0211	432826.5000 ± 286.9872	0.0%	0.0099 ± 0.0043	3053.5000 ± 1180.2092	0.6327 ± 0.2000	183000.0000 ± 59615.4342	100.0%	91/94
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	1.58308 ± 0.1846	426325.3000 ± 997.9812	0.0%	0.0116 ± 0.0070	2836.6000 ± 1404.7186	0.5289 ± 0.2234	136100.0000 ± 49902.8055	100.0%	91/95
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	1.47583 ± 0.0345	427504.5000 ± 1672.9297	0.0%	0.0092 ± 0.0057	2317.6000 ± 973.4065	0.6362 ± 0.3604	184333.3333 ± 105651.4185	60.0%	91/413
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	2.26246 ± 0.0669	663258.0000 ± 1233.7467	0.0%	0.0173 ± 0.0077	5240.3000 ± 2291.8347	0.6606 ± 0.1448	196500.0000 ± 38933.9184	100.0%	91/95
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	2.17321 ± 0.0559	646577.3000 ± 1235.0955	0.0%	0.0156 ± 0.0066	4726.7000 ± 1790.8730	0.7594 ± 0.4649	229800.0000 ± 142064.6332	100.0%	91/96
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	2.20964 ± 0.0344	642264.3000 ± 1757.2241	0.0%	0.0134 ± 0.0083	3833.0000 ± 2582.2064	1.0991 ± 0.7577	317857.1429 ± 215001.2814	70.0%	91/93
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	2.19196 ± 0.0347	640818.9000 ± 2028.4827	0.0%	0.0168 ± 0.0107	4567.4000 ± 2815.3857	1.5732 ± 0.0000	468000.0000 ± 0.0000	10.0%	91/541
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	4.43949 ± 0.1305	1345399.1000 ± 4435.1527	0.0%	0.0272 ± 0.0142	7816.0000 ± 3692.5878	0.7407 ± 0.1363	220800.0000 ± 30514.9144	100.0%	91/94
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	4.25523 ± 0.0747	1336149.6000 ± 2826.8964	0.0%	0.0279 ± 0.0157	8969.9000 ± 5389.6546	0.7795 ± 0.3274	246800.0000 ± 99628.1085	100.0%	91/94
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	4.24001 ± 0.0731	1334894.4000 ± 4073.3894	0.0%	0.0310 ± 0.0157	10006.7000 ± 4786.0015	1.5722 ± 0.9591	491800.0000 ± 297310.2084	100.0%	91/96
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	4.28582 ± 0.0528	1335586.1000 ± 4076.2827	0.0%	0.0261 ± 0.0060	8509.5000 ± 1956.4541	2.4463 ± 1.4139	763000.0000 ± 424209.8537	30.0%	91/94
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	10.90180 ± 0.3767	3702959.7000 ± 18502.6419	0.0%	0.0586 ± 0.0249	20067.2000 ± 8975.6377	0.9018 ± 0.1545	307800.0000 ± 46807.6917	100.0%	91/95
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	10.73192 ± 0.0745	3719210.1000 ± 25247.8884	0.0%	0.0323 ± 0.0182	11336.6000 ± 6072.4067	1.0070 ± 0.3160	353100.0000 ± 109684.5021	100.0%	91/95
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	10.73638 ± 0.1441	3716656.7000 ± 17669.6110	0.0%	0.0499 ± 0.0196	17346.3000 ± 6210.4451	2.4042 ± 2.7515	833100.0000 ± 949126.5932	100.0%	91/644
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	10.77606 ± 0.1187	3737649.6000 ± 22348.5651	0.0%	0.0415 ± 0.0264	14317.1000 ± 8608.5730	4.0375 ± 2.2799	1406777.7778 ± 798277.6985	90.0%	91/508
Q-learning ($\rho = 0.5, \epsilon = 1$)	25.58737 ± 0.4112	8949239.4000 ± 8155.9487	0.0%	0.2668 ± 0.2094	91797.6000 ± 74593.1546	7.5097 ± 1.8759	2623700.0000 ± 669899.1043	100.0%	91/95
Q-learning ($\rho = 0.7, \epsilon = 1$)	25.40030 ± 0.2116	8952399.5000 ± 7496.6479	0.0%	0.3933 ± 0.4682	139007.4000 ± 164452.4483	4.8250 ± 0.9852	1681300.0000 ± 339919.4169	100.0%	91/94
Q-learning ($\rho = 0.9, \epsilon = 1$)	25.38862 ± 0.2153	8947227.8000 ± 6037.7618	0.0%	0.4111 ± 0.2980	146740.3000 ± 105196.9891	4.3187 ± 1.0141	1525900.0000 ± 358173.8265	100.0%	91/867
Q-learning ($\rho = 0.99, \epsilon = 1$)	25.44583 ± 0.1526	8949491.3000 ± 7905.9290	0.0%	0.3717 ± 0.4225	131063.6000 ± 147018.7102	6.5359 ± 1.8032	2301600.0000 ± 641303.8593	100.0%	91/94
Async Value Iteration	0.13077 ± 0.0088	153850.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	91/96
Value Iteration	0.22881 ± 0.0077	276930.0000 ± 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 ρ 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 ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	1.21094 ± 0.0463	344046.1000 ± 102.2814	0.0%	0.0084 ± 0.0012	2451.4000 ± 110.7729	0.4106 ± 0.0192	117600.0000 ± 1685.2300	100.0%	91/91
Q-learning ($\rho = 0.7, \epsilon = 0$)	1.10425 ± 0.0174	323271.1000 ± 485.7095	0.0%	0.0112 ± 0.0019	3345.3000 ± 552.7379	0.2849 ± 0.0118	83300.0000 ± 2002.4984	100.0%	91/94
Q-learning ($\rho = 0.9, \epsilon = 0$)	1.10495 ± 0.0342	314963.3000 ± 2060.9715	0.0%	0.0098 ± 0.0019	2861.3000 ± 469.4173	0.2285 ± 0.0140	64300.0000 ± 1100.0000	100.0%	91/3322
Q-learning ($\rho = 0.99, \epsilon = 0$)	1.13159 ± 0.0396	312640.0000 ± 1307.9674	0.0%	0.0112 ± 0.0040	2865.7000 ± 362.5336	0.2073 ± 0.0200	57800.0000 ± 871.7798	100.0%	91/11470
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	1.55639 ± 0.0474	444768.9000 ± 672.7366	0.0%	0.0133 ± 0.0072	3785.6000 ± 2061.8938	0.4170 ± 0.0217	128900.0000 ± 2022.3748	100.0%	91/91
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	1.47309 ± 0.0541	425557.3000 ± 477.7305	0.0%	0.0074 ± 0.0016	2284.2000 ± 457.0936	0.3191 ± 0.0183	93600.0000 ± 1356.4660	100.0%	91/91
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	1.45958 ± 0.0953	415118.0000 ± 1083.8050	0.0%	0.0115 ± 0.0070	3451.3000 ± 2222.0212	0.2585 ± 0.0230	74000.0000 ± 894.4272	100.0%	91/93
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	1.43543 ± 0.0305	412530.3000 ± 690.9857	0.0%	0.0119 ± 0.0070	3408.6000 ± 1994.6657	0.2363 ± 0.0160	67400.0000 ± 1356.4660	100.0%	91/91
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	2.24574 ± 0.0475	654971.4000 ± 1405.2715	0.0%	0.0180 ± 0.0106	4935.3000 ± 2842.1437	0.4980 ± 0.0205	145700.0000 ± 1900.0000	100.0%	91/91
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	2.17102 ± 0.0289	640450.2000 ± 1371.2788	0.0%	0.0158 ± 0.0057	4673.3000 ± 1744.4904	0.3724 ± 0.0069	111100.0000 ± 1374.7727	100.0%	91/92
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	2.16785 ± 0.0543	631810.6000 ± 1227.3894	0.0%	0.0175 ± 0.0071	5026.0000 ± 1940.3128	0.3351 ± 0.0172	90500.0000 ± 1204.1595	100.0%	91/91
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	2.11225 ± 0.0409	627377.9000 ± 1655.9167	0.0%	0.0151 ± 0.0062	4572.4000 ± 1658.9468	0.2787 ± 0.0118	82900.0000 ± 1300.0000	100.0%	91/93
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	4.47757 ± 0.1016	1345727.5000 ± 3977.2152	0.0%	0.0328 ± 0.0141	9890.0000 ± 4365.4134	0.6339 ± 0.0323	188600.0000 ± 3720.2150	100.0%	91/91
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	4.43670 ± 0.1067	1333023.2000 ± 4511.8552	0.0%	0.0229 ± 0.0132	7162.9000 ± 4239.4070	0.4808 ± 0.0169	146600.0000 ± 3104.8349	100.0%	91/93
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	4.52178 ± 0.1643	1330597.8000 ± 3633.1147	0.0%	0.0263 ± 0.0160	7498.3000 ± 4079.6333	0.4121 ± 0.0219	123200.0000 ± 1661.3248	100.0%	91/91
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	4.48334 ± 0.0950	1321232.6000 ± 4239.8322	0.0%	0.0246 ± 0.0126	7053.3000 ± 3504.0261	0.3902 ± 0.0200	114900.0000 ± 2507.9872	100.0%	91/91
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	2.81375 ± 0.6228	807321.3000 ± 175701.3814	100.0%	0.0539 ± 0.0272	16241.7000 ± 8263.9820	0.8684 ± 0.0333	254400.0000 ± 7459.2225	100.0%	91/93
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	5.23973 ± 3.3760	1490325.2000 ± 955458.1602	90.0%	0.0490 ± 0.0256	14057.6000 ± 7212.0513	0.7160 ± 0.0281	204600.0000 ± 10041.9122	100.0%	91/91
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	10.40519 ± 2.8531	3044238.7000 ± 808067.4336	50.0%	0.0639 ± 0.0330	18087.4000 ± 8803.0574	0.6247 ± 0.0540	182200.0000 ± 9378.6993	100.0%	91/91
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	12.12238 ± 0.1392	3727107.8000 ± 25514.4807	0.0%	0.0660 ± 0.0335	20051.8000 ± 10664.1973	0.5289 ± 0.0282	163800.0000 ± 8852.1184	100.0%	91/92
Q-learning ($\rho = 0.5, \epsilon = 1$)	12.78287 ± 3.3036	3996847.8000 ± 1030511.9982	100.0%	0.4339 ± 0.3682	137499.1000 ± 117908.5560	7.9335 ± 1.4208	2479600.0000 ± 452224.9883	100.0%	91/93
Q-learning ($\rho = 0.7, \epsilon = 1$)	10.58299 ± 2.2827	3299660.7000 ± 713682.8484	100.0%	0.4017 ± 0.4531	122460.6000 ± 140043.9408	7.4290 ± 0.9222	2317200.0000 ± 289360.6055	100.0%	91/91
Q-learning ($\rho = 0.9, \epsilon = 1$)	8.27295 ± 1.7336	2561736.4000 ± 526837.9684	100.0%	0.3989 ± 0.2493	126615.4000 ± 82858.7311	5.3974 ± 1.8055	1678400.0000 ± 563602.3776	100.0%	91/92
Q-learning ($\rho = 0.99, \epsilon = 1$)	10.17175 ± 4.4533	3250319.2000 ± 1461986.1491	100.0%	0.2085 ± 0.2196	66737.3000 ± 70059.1983	5.4702 ± 2.0360	1745200.0000 ± 650229.7748	100.0%	91/93
Async Value Iteration	0.12174 ± 0.0075	130320.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	91/91
Value Iteration	0.22499 ± 0.0142	238920.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	91/93

Table 49. Comparing performance of Q-learning as we change the learning rate ρ in stochastic problem with $\gamma = 0.99$ (Problem 13). Two versions of value iteration are also appended for contrast

Algorithm (Problem 13) ($\gamma = 0.99$)	Runtime (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.41708 ± 0.0074	130207.0000 ± 114.5111	0.0%	0.0054 ± 0.0008	1592.3000 ± 215.2181	0.1329 ± 0.0121	41800.0000 ± 1886.7962	100.0%	33/35
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.41080 ± 0.0134	126811.0000 ± 1870.6015	0.0%	0.0048 ± 0.0014	1537.0000 ± 238.6487	0.1109 ± 0.0246	34444.4444 ± 7917.7874	90.0%	33/49
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.41992 ± 0.0111	130207.7000 ± 2948.2521	0.0%	0.0048 ± 0.0012	1533.3000 ± 297.4283	0.1099 ± 0.0487	34666.6667 ± 14204.8505	90.0%	33/39
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.42076 ± 0.0123	131368.2000 ± 2567.8135	0.0%	0.0046 ± 0.0008	1552.9000 ± 252.3705	0.1010 ± 0.0089	32666.6667 ± 2357.0226	30.0%	33/1624
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.54763 ± 0.0160	173912.5000 ± 487.3506	0.0%	0.0048 ± 0.0015	1512.5000 ± 238.4438	0.1504 ± 0.0287	48500.0000 ± 7826.2379	100.0%	33/36
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.52066 ± 0.0093	165434.7000 ± 493.1110	0.0%	0.0051 ± 0.0010	1505.3000 ± 227.6686	0.1328 ± 0.0261	42300.0000 ± 9121.9515	100.0%	33/35
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	0.51072 ± 0.0110	162109.3000 ± 907.4319	0.0%	0.0047 ± 0.0014	1530.4000 ± 374.2005	0.1623 ± 0.0894	52000.0000 ± 29492.3719	100.0%	33/39
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	0.50824 ± 0.0139	160895.5000 ± 779.3799	0.0%	0.0053 ± 0.0008	1665.0000 ± 273.3050	nan ± nan	nan ± nan	0.0%	33/1517
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.81460 ± 0.0118	265634.6000 ± 675.8025	0.0%	0.0061 ± 0.0018	2090.8000 ± 543.4166	0.1639 ± 0.0126	53900.0000 ± 3330.1652	100.0%	33/35
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.79613 ± 0.0191	257595.1000 ± 1356.6870	0.0%	0.0057 ± 0.0013	1858.4000 ± 362.2282	0.1466 ± 0.0282	48000.0000 ± 7758.8659	100.0%	33/35
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	0.79041 ± 0.0325	253944.2000 ± 1311.6118	0.0%	0.0057 ± 0.0021	1908.2000 ± 647.3598	0.2370 ± 0.1439	74700.0000 ± 39709.0670	100.0%	33/35
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.78437 ± 0.0100	253555.1000 ± 802.9506	0.0%	0.0055 ± 0.0009	1853.8000 ± 265.7265	nan ± nan	nan ± nan	0.0%	33/36
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	1.71005 ± 0.0112	572838.0000 ± 2249.6028	0.0%	0.0100 ± 0.0032	3297.9000 ± 922.6598	0.2122 ± 0.0196	73400.0000 ± 6545.2273	100.0%	33/34
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	1.71742 ± 0.0165	570048.0000 ± 4754.1227	0.0%	0.0095 ± 0.0028	3087.4000 ± 902.2534	0.1788 ± 0.0213	60300.0000 ± 7603.2888	100.0%	33/34
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	1.71999 ± 0.0195	568907.4000 ± 2242.6088	0.0%	0.0091 ± 0.0030	2924.0000 ± 1067.4881	0.2768 ± 0.0793	93800.0000 ± 26006.9222	100.0%	33/35
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	1.73093 ± 0.0150	573088.3000 ± 4372.8894	0.0%	0.0094 ± 0.0035	3057.7000 ± 1176.1283	1.2025 ± 0.0000	397000.0000 ± 0.0000	10.0%	33/35
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	4.90308 ± 0.0998	1648811.3000 ± 21705.7912	0.0%	0.0113 ± 0.0057	3674.2000 ± 1768.6525	0.3200 ± 0.0605	109400.0000 ± 20085.8159	100.0%	33/36
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	4.92352 ± 0.0782	1662491.4000 ± 17743.7073	0.0%	0.0061 ± 0.0031	2356.7000 ± 1105.4382	0.2610 ± 0.1042	92100.0000 ± 36911.9222	100.0%	33/36
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	5.01483 ± 0.0545	1695821.1000 ± 14854.3833	0.0%	0.0118 ± 0.0063	3860.2000 ± 2004.5013	0.3561 ± 0.2383	123100.0000 ± 81007.9626	100.0%	33/35
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	5.17860 ± 0.0722	1749412.8000 ± 17880.0877	0.0%	0.0095 ± 0.0063	3261.4000 ± 1945.5436	nan ± nan	nan ± nan	0.0%	33/35
Q-learning ($\rho = 0.5, \epsilon = 1$)	20.19759 ± 0.2087	6935698.3000 ± 59030.1047	0.0%	0.0140 ± 0.0123	4713.7000 ± 4234.9662	0.7662 ± 0.0756	266500.0000 ± 23580.7125	100.0%	33/35
Q-learning ($\rho = 0.7, \epsilon = 1$)	20.25416 ± 0.2529	6959126.6000 ± 60221.3020	0.0%	0.0072 ± 0.0051	2533.0000 ± 1668.3765	0.6349 ± 0.1335	219100.0000 ± 44650.7559	100.0%	33/35
Q-learning ($\rho = 0.9, \epsilon = 1$)	20.19583 ± 0.1558	6968889.3000 ± 54469.5879	0.0%	0.0139 ± 0.0092	4875.1000 ± 2973.4560	0.6057 ± 0.1479	212100.0000 ± 49383.0943	100.0%	33/34
Q-learning ($\rho = 0.99, \epsilon = 1$)	20.34780 ± 0.1439	6978691.0000 ± 50001.2043	0.0%	0.0141 ± 0.0133	4624.0000 ± 4342.0776	nan ± nan	nan ± nan	0.0%	33/34
Async Value Iteration	0.09685 ± 0.0033	137088.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	33/33
Value Iteration	0.14306 ± 0.0030	204288.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	33/34

Table 50. Comparing performance of Q-learning as we change the learning rate ρ 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 ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.42100 ± 0.0143	127752.2000 ± 108.4627	0.0%	0.0047 ± 0.0008	1477.1000 ± 76.6022	0.1265 ± 0.0068	39100.0000 ± 300.0000	100.0%	33/33
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.38638 ± 0.0109	119281.5000 ± 378.1992	0.0%	0.0042 ± 0.0007	1313.6000 ± 65.3578	0.0914 ± 0.0029	28400.0000 ± 489.8979	100.0%	33/33
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.37736 ± 0.0127	117003.6000 ± 1313.0820	0.0%	0.0044 ± 0.0008	1309.9000 ± 33.8274	0.0727 ± 0.0038	22800.0000 ± 600.0000	100.0%	33/35
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.36804 ± 0.0097	115082.0000 ± 1690.8937	0.0%	0.0042 ± 0.0007	1369.4000 ± 153.5579	0.0646 ± 0.0044	20800.0000 ± 871.7798	100.0%	33/35
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.53299 ± 0.0136	170583.4000 ± 373.3465	0.0%	0.0051 ± 0.0016	1629.0000 ± 322.1391	0.1338 ± 0.0029	42900.0000 ± 300.0000	100.0%	33/33
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.50840 ± 0.0103	161753.3000 ± 484.6933	0.0%	0.0049 ± 0.0008	1532.1000 ± 224.2046	0.1000 ± 0.0029	32100.0000 ± 300.0000	100.0%	33/33
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	0.49029 ± 0.0129	156851.3000 ± 651.1839	0.0%	0.0053 ± 0.0012	1599.5000 ± 332.4681	0.0806 ± 0.0070	25400.0000 ± 489.8979	100.0%	33/33
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	0.48949 ± 0.0121	155549.2000 ± 427.4459	0.0%	0.0045 ± 0.0011	1478.0000 ± 309.7082	0.0749 ± 0.0037	24100.0000 ± 538.5165	100.0%	33/34
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.80326 ± 0.0161	259210.8000 ± 1509.3220	0.0%	0.0062 ± 0.0016	2006.1000 ± 345.7610	0.1541 ± 0.0074	49700.0000 ± 640.3124	100.0%	33/33
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.77900 ± 0.0185	252298.5000 ± 1926.2094	0.0%	0.0068 ± 0.0020	2069.7000 ± 648.0762	0.1176 ± 0.0058	37800.0000 ± 400.0000	100.0%	33/33
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	0.76413 ± 0.0141	248537.1000 ± 1056.6522	0.0%	0.0057 ± 0.0014	2003.1000 ± 303.4526	0.0934 ± 0.0028	31000.0000 ± 0.0000	100.0%	33/33
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.77291 ± 0.0190	247948.2000 ± 1162.4237	0.0%	0.0061 ± 0.0019	1859.1000 ± 547.9541	0.0910 ± 0.0072	29000.0000 ± 447.2136	100.0%	33/33
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	1.71760 ± 0.0251	573371.0000 ± 6643.6816	0.0%	0.0083 ± 0.0020	2860.5000 ± 685.0938	0.1896 ± 0.0083	64300.0000 ± 1100.0000	100.0%	33/33
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	1.76432 ± 0.0306	585333.7000 ± 5937.4857	0.0%	0.0070 ± 0.0021	2409.1000 ± 629.4284	0.1503 ± 0.0055	50700.0000 ± 1791.6473	100.0%	33/33
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	1.79295 ± 0.0416	592897.9000 ± 6914.9360	0.0%	0.0091 ± 0.0031	2858.0000 ± 817.7569	0.1283 ± 0.0073	42600.0000 ± 800.0000	100.0%	33/33
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	1.74500 ± 0.0173	587144.2000 ± 4339.3984	0.0%	0.0064 ± 0.0016	2172.8000 ± 602.0843	0.1150 ± 0.0054	40100.0000 ± 1513.2746	100.0%	33/34
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	0.55173 ± 0.0504	189588.4000 ± 17275.8681	100.0%	0.0103 ± 0.0053	3651.0000 ± 1959.4914	0.2617 ± 0.0102	89500.0000 ± 3008.3218	100.0%	33/33
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	0.45882 ± 0.0514	159215.7000 ± 17723.4830	100.0%	0.0104 ± 0.0053	3641.1000 ± 1702.7189	0.2070 ± 0.0162	72800.0000 ± 4354.3082	100.0%	33/33
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	0.42473 ± 0.0747	142218.5000 ± 25592.6479	100.0%	0.0127 ± 0.0065	4168.0000 ± 2153.3371	0.1819 ± 0.0126	60800.0000 ± 3280.2439	100.0%	33/33
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	0.35403 ± 0.0764	119643.8000 ± 25616.2213	100.0%	0.0107 ± 0.0045	3898.7000 ± 1632.3342	0.1696 ± 0.0126	57500.0000 ± 3383.7849	100.0%	33/33
Q-learning ($\rho = 0.5, \epsilon = 1$)	1.14662 ± 0.2058	400989.4000 ± 71658.6692	100.0%	0.0146 ± 0.0141	5338.0000 ± 5032.0724	0.7876 ± 0.1240	275900.0000 ± 44468.9780	100.0%	33/35
Q-learning ($\rho = 0.7, \epsilon = 1$)	0.99436 ± 0.1848	346500.0000 ± 61741.7089	100.0%	0.0128 ± 0.0110	4797.6000 ± 4424.0289	0.6524 ± 0.0992	227700.0000 ± 32972.8676	100.0%	33/33
Q-learning ($\rho = 0.9, \epsilon = 1$)	0.76700 ± 0.1771	265314.4000 ± 60957.6632	100.0%	0.0081 ± 0.0051	2921.1000 ± 1713.3967	0.4927 ± 0.0841	170200.0000 ± 28424.6372	100.0%	33/33
Q-learning ($\rho = 0.99, \epsilon = 1$)	0.62492 ± 0.1279	218100.4000 ± 43522.9246	100.0%	0.0106 ± 0.0065	3915.7000 ± 2612.0314	0.4177 ± 0.0718	145400.0000 ± 26131.2074	100.0%	33/33
Async Value Iteration	0.07626 ± 0.0034	112896.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	33/33
Value Iteration	0.11517 ± 0.0051	163968.0000 ± 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 ρ 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 ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.30970 ± 0.0091	96328.6000 ± 176.8594	0.0%	0.0040 ± 0.0009	1252.6000 ± 265.3127	0.1077 ± 0.0078	34800.0000 ± 1833.0303	100.0%	24/32
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.29002 ± 0.0048	91828.5000 ± 931.1332	0.0%	0.0049 ± 0.0014	1435.9000 ± 234.8917	0.0769 ± 0.0081	24600.0000 ± 2107.1308	100.0%	24/28
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.30375 ± 0.0170	93267.8000 ± 1501.4346	0.0%	0.0046 ± 0.0012	1352.2000 ± 270.5168	0.0645 ± 0.0080	20000.0000 ± 2529.8221	50.0%	24/28
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.31794 ± 0.0239	94536.9000 ± 1658.8502	0.0%	0.0044 ± 0.0008	1390.3000 ± 218.5599	0.1913 ± 0.0751	57666.6667 ± 20717.6790	60.0%	24/239
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.40410 ± 0.0155	126466.8000 ± 334.4978	0.0%	0.0063 ± 0.0008	1860.4000 ± 358.0489	0.1197 ± 0.0136	38000.0000 ± 3768.2887	100.0%	24/26
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.38891 ± 0.0115	120149.8000 ± 290.1523	0.0%	0.0050 ± 0.0014	1568.2000 ± 362.5719	0.1007 ± 0.0166	32100.0000 ± 5166.2365	100.0%	24/25
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	0.36089 ± 0.0113	117124.6000 ± 603.3306	0.0%	0.0047 ± 0.0013	1479.7000 ± 294.5665	0.1410 ± 0.0746	46300.0000 ± 25772.2719	100.0%	24/26
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	0.38105 ± 0.0138	116433.2000 ± 451.0545	0.0%	0.0063 ± 0.0015	1825.3000 ± 231.3846	0.1326 ± 0.0000	42000.0000 ± 0.0000	10.0%	24/25
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.59012 ± 0.0191	190592.8000 ± 240.0416	0.0%	0.0071 ± 0.0019	2195.3000 ± 402.9462	0.1289 ± 0.0045	41400.0000 ± 1496.6630	100.0%	24/26
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.56367 ± 0.0122	184919.0000 ± 561.0007	0.0%	0.0065 ± 0.0014	2107.4000 ± 306.1615	0.1137 ± 0.0268	37000.0000 ± 7549.8344	100.0%	24/26
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	0.55118 ± 0.0081	182080.0000 ± 618.4603	0.0%	0.0059 ± 0.0018	1922.5000 ± 459.6375	0.1083 ± 0.0255	35700.0000 ± 8343.2608	100.0%	24/27
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.55182 ± 0.0097	181445.6000 ± 872.0140	0.0%	0.0065 ± 0.0014	2058.4000 ± 397.7271	0.2859 ± 0.2007	96000.0000 ± 67000.0000	20.0%	24/24
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	1.18849 ± 0.0185	397074.7000 ± 1116.4341	0.0%	0.0067 ± 0.0024	2207.1000 ± 846.5800	0.1594 ± 0.0095	54300.0000 ± 2968.1644	100.0%	24/25
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	1.17601 ± 0.0220	392153.6000 ± 2179.3055	0.0%	0.0079 ± 0.0032	2682.2000 ± 1006.5925	0.1436 ± 0.0217	48500.0000 ± 5852.3500	100.0%	24/25
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	1.19923 ± 0.0324	390143.3000 ± 2422.7454	0.0%	0.0067 ± 0.0021	2280.6000 ± 621.1773	0.1828 ± 0.0843	61200.0000 ± 28099.1103	100.0%	24/26
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	1.16542 ± 0.0128	394794.1000 ± 2573.2271	0.0%	0.0088 ± 0.0036	2861.0000 ± 829.6588	nan ± nan	nan ± nan	0.0%	24/26
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	3.08581 ± 0.0616	1040344.3000 ± 9700.2659	0.0%	0.0073 ± 0.0054	2100.6000 ± 1351.0389	0.2479 ± 0.0600	81300.0000 ± 16174.3624	100.0%	24/27
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	3.04500 ± 0.0401	1042682.5000 ± 8516.5642	0.0%	0.0061 ± 0.0034	2100.2000 ± 1257.2713	0.1948 ± 0.0451	66100.0000 ± 13794.5641	100.0%	24/24
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	3.07272 ± 0.0809	1043268.7000 ± 13685.6012	0.0%	0.0100 ± 0.0045	3357.6000 ± 1384.5272	0.2213 ± 0.1023	76000.0000 ± 35547.1518	100.0%	24/26
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	3.15766 ± 0.0590	1075696.7000 ± 10863.5693	0.0%	0.0082 ± 0.0058	2911.9000 ± 1920.4674	nan ± nan	nan ± nan	0.0%	24/27
Q-learning ($\rho = 0.5, \epsilon = 1$)	13.48848 ± 0.2029	4627804.4000 ± 42637.9797	0.0%	0.0060 ± 0.0042	1833.1000 ± 1449.0891	0.3290 ± 0.0742	112000.0000 ± 21090.2821	100.0%	24/25
Q-learning ($\rho = 0.7, \epsilon = 1$)	13.30236 ± 0.2259	4594436.8000 ± 55299.0742	0.0%	0.0054 ± 0.0029	2101.6000 ± 1352.8404	0.2844 ± 0.0377	100400.0000 ± 13792.7517	100.0%	24/26
Q-learning ($\rho = 0.9, \epsilon = 1$)	13.40463 ± 0.1464	4644797.6000 ± 44227.3172	0.0%	0.0060 ± 0.0046	2046.6000 ± 1783.7949	0.3926 ± 0.1898	136600.0000 ± 62220.8968	100.0%	24/25
Q-learning ($\rho = 0.99, \epsilon = 1$)	13.34453 ± 0.1413	4619388.6000 ± 24355.0382	0.0%	0.0042 ± 0.0020	1506.4000 ± 885.7109	nan ± nan	nan ± nan	0.0%	24/26
Async Value Iteration	0.10068 ± 0.0071	129688.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	24/26
Value Iteration	0.13079 ± 0.0037	167098.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	24/24

Table 52. Comparing performance of Q-learning as we change the learning rate ρ in stochastic problem with $\gamma = 0.999$ (Problem 14). Two versions of value iteration are also appended for contrast

Algorithm (Problem 14) ($\gamma = 0.999$)	Runtime (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.30509 ± 0.0127	94365.1000 ± 115.9418	0.0%	0.0048 ± 0.0008	1555.4000 ± 153.3651	0.1033 ± 0.0033	32000.0000 ± 0.0000	100.0%	24/24
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.27763 ± 0.0110	87702.9000 ± 163.8111	0.0%	0.0040 ± 0.0008	1311.1000 ± 38.2164	0.0720 ± 0.0036	23200.0000 ± 400.0000	100.0%	24/24
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.27524 ± 0.0138	85869.9000 ± 1426.7016	0.0%	0.0042 ± 0.0011	1297.5000 ± 41.6371	0.0582 ± 0.0070	18000.0000 ± 0.0000	100.0%	24/26
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.26893 ± 0.0069	85089.9000 ± 1288.8214	0.0%	0.0042 ± 0.0005	1305.5000 ± 49.4131	0.0529 ± 0.0018	17000.0000 ± 0.0000	100.0%	24/26
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.38798 ± 0.0101	124627.3000 ± 316.6519	0.0%	0.0060 ± 0.0014	2001.7000 ± 475.6738	0.1090 ± 0.0040	35100.0000 ± 538.5165	100.0%	24/26
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.37066 ± 0.0159	117827.2000 ± 399.2553	0.0%	0.0049 ± 0.0011	1600.8000 ± 273.2595	0.0810 ± 0.0037	26100.0000 ± 830.6624	100.0%	24/24
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	0.36039 ± 0.0103	114754.0000 ± 408.0216	0.0%	0.0051 ± 0.0020	1632.0000 ± 420.1623	0.0643 ± 0.0020	20800.0000 ± 600.0000	100.0%	24/24
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	0.36082 ± 0.0145	114243.6000 ± 262.6596	0.0%	0.0047 ± 0.0011	1510.6000 ± 255.3500	0.0609 ± 0.0037	19200.0000 ± 400.0000	100.0%	24/24
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.56919 ± 0.0057	187718.5000 ± 1252.0193	0.0%	0.0062 ± 0.0014	1978.8000 ± 378.8931	0.1217 ± 0.0038	40400.0000 ± 663.3250	100.0%	24/24
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.55891 ± 0.0108	182511.2000 ± 991.1749	0.0%	0.0057 ± 0.0019	1979.6000 ± 470.0022	0.0922 ± 0.0076	30500.0000 ± 806.2258	100.0%	24/24
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	0.54539 ± 0.0124	180800.3000 ± 1082.4676	0.0%	0.0072 ± 0.0011	2337.0000 ± 276.2238	0.0740 ± 0.0043	24700.0000 ± 640.3124	100.0%	24/24
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.55363 ± 0.0083	181992.7000 ± 734.9201	0.0%	0.0068 ± 0.0012	2221.6000 ± 380.8276	0.0698 ± 0.0025	23400.0000 ± 489.8979	100.0%	24/24
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	1.18523 ± 0.0191	399329.6000 ± 4719.2824	0.0%	0.0083 ± 0.0028	3098.4000 ± 924.0941	0.1487 ± 0.0047	50500.0000 ± 1204.1595	100.0%	24/24
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	1.20939 ± 0.0309	402571.2000 ± 2161.2340	0.0%	0.0074 ± 0.0032	2653.8000 ± 1258.6876	0.1212 ± 0.0071	40600.0000 ± 1200.0000	100.0%	24/25
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	1.21022 ± 0.0218	408583.6000 ± 3412.3200	0.0%	0.0070 ± 0.0033	2565.7000 ± 1113.2647	0.0984 ± 0.0052	33600.0000 ± 1200.0000	100.0%	24/24
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	1.23702 ± 0.0155	414506.0000 ± 3230.7090	0.0%	0.0068 ± 0.0029	2220.9000 ± 735.6789	0.0947 ± 0.0063	31500.0000 ± 1024.6951	100.0%	24/24
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	1.24375 ± 0.2250	428780.2000 ± 75699.6203	100.0%	0.0060 ± 0.0045	2044.9000 ± 1605.6018	0.1940 ± 0.0089	68400.0000 ± 2244.9944	100.0%	24/24
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	0.77661 ± 0.1855	263318.7000 ± 49856.2033	100.0%	0.0068 ± 0.0042	2432.7000 ± 1644.0220	0.1567 ± 0.0110	54300.0000 ± 3579.1060	100.0%	24/24
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	0.68355 ± 0.1465	237633.9000 ± 52024.9411	100.0%	0.0074 ± 0.0029	2495.0000 ± 896.0204	0.1376 ± 0.0075	46400.0000 ± 2537.7155	100.0%	24/24
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	0.44536 ± 0.0739	154420.1000 ± 24849.9184	100.0%	0.0060 ± 0.0036	2145.6000 ± 1170.3625	0.1210 ± 0.0092	42000.0000 ± 2489.9799	100.0%	24/24
Q-learning ($\rho = 0.5, \epsilon = 1$)	0.49448 ± 0.0793	173748.1000 ± 27615.5621	100.0%	0.0045 ± 0.0028	1543.5000 ± 1051.5644	0.2947 ± 0.0408	104400.0000 ± 11499.5652	100.0%	24/24
Q-learning ($\rho = 0.7, \epsilon = 1$)	0.39339 ± 0.0187	139320.9000 ± 9283.7602	100.0%	0.0069 ± 0.0049	2119.8000 ± 1404.8083	0.2316 ± 0.0306	82100.0000 ± 10300.0000	100.0%	24/24
Q-learning ($\rho = 0.9, \epsilon = 1$)	0.33914 ± 0.0224	121377.5000 ± 7441.4884	100.0%	0.0039 ± 0.0029	1332.8000 ± 1053.2151	0.1998 ± 0.0115	70900.0000 ± 4482.1870	100.0%	24/24
Q-learning ($\rho = 0.99, \epsilon = 1$)	0.35365 ± 0.0521	124311.0000 ± 16302.3570	100.0%	0.0070 ± 0.0048	2020.3000 ± 1504.0803	0.1803 ± 0.0203	63700.0000 ± 8100.0000	100.0%	24/24
Async Value Iteration	0.08128 ± 0.0039	107242.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	24/24
Value Iteration	0.10199 ± 0.0044	132182.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	24/25

Table 53. Comparing performance of Q-learning as we change the learning rate ρ 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 ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.28852 ± 0.0117	89028.3000 ± 184.7880	0.0%	0.0048 ± 0.0007	1428.2000 ± 129.3706	0.1241 ± 0.0113	38600.0000 ± 3261.9013	100.0%	22/24
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.26617 ± 0.0100	83249.1000 ± 321.7780	0.0%	0.0053 ± 0.0012	1495.4000 ± 203.8942	0.0940 ± 0.0080	29900.0000 ± 2426.9322	100.0%	22/25
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.26301 ± 0.0066	81721.2000 ± 1224.0630	0.0%	0.0055 ± 0.0008	1565.3000 ± 179.0643	0.0876 ± 0.0136	27000.0000 ± 4404.5431	100.0%	22/26
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.26566 ± 0.0046	81982.9000 ± 1051.3302	0.0%	0.0049 ± 0.0012	1413.1000 ± 133.2039	0.1278 ± 0.0369	40000.0000 ± 11175.3697	90.0%	22/26
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.36004 ± 0.0117	113794.9000 ± 352.9361	0.0%	0.0058 ± 0.0018	1701.5000 ± 314.5973	0.1777 ± 0.0603	56000.0000 ± 18487.8338	100.0%	22/24
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.34671 ± 0.0133	107591.1000 ± 295.7453	0.0%	0.0050 ± 0.0013	1596.3000 ± 309.2633	0.2252 ± 0.0689	70333.3333 ± 19595.9179	90.0%	22/23
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	0.33285 ± 0.0086	104282.5000 ± 301.7327	0.0%	0.0052 ± 0.0010	1704.4000 ± 263.9326	0.2373 ± 0.0904	73750.0000 ± 28145.8256	40.0%	22/23
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	0.32981 ± 0.0066	103699.3000 ± 399.7337	0.0%	0.0055 ± 0.0020	1501.7000 ± 434.1223	nan ± nan	nan ± nan	0.0%	22/25
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.51373 ± 0.0124	165242.8000 ± 657.2284	0.0%	0.0073 ± 0.0016	2317.4000 ± 602.3433	0.2117 ± 0.0821	69200.0000 ± 26794.7756	100.0%	22/22
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.49143 ± 0.0062	159776.7000 ± 586.9612	0.0%	0.0055 ± 0.0015	1763.9000 ± 409.8842	0.1953 ± 0.0493	62900.0000 ± 15712.7337	100.0%	22/22
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	0.48423 ± 0.0164	156980.0000 ± 740.5661	0.0%	0.0057 ± 0.0017	1920.3000 ± 280.7729	0.2244 ± 0.0650	75250.0000 ± 23413.4043	40.0%	22/23
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.48535 ± 0.0111	156709.0000 ± 711.7792	0.0%	0.0045 ± 0.0011	1662.9000 ± 376.2540	nan ± nan	nan ± nan	0.0%	22/23
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	0.98409 ± 0.0171	329468.7000 ± 2534.9306	0.0%	0.0076 ± 0.0021	2541.0000 ± 668.8504	0.2437 ± 0.0631	83200.0000 ± 22076.2316	100.0%	22/25
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	0.98745 ± 0.0149	326236.2000 ± 2648.0536	0.0%	0.0059 ± 0.0022	2039.6000 ± 651.3889	0.3675 ± 0.2205	122700.0000 ± 70218.3025	100.0%	22/25
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	0.97015 ± 0.0175	324972.7000 ± 1958.8229	0.0%	0.0055 ± 0.0025	1835.6000 ± 908.8369	0.4076 ± 0.1862	137500.0000 ± 60864.6038	80.0%	22/23
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	0.96860 ± 0.0149	324020.9000 ± 2793.0568	0.0%	0.0063 ± 0.0019	2181.8000 ± 593.4130	nan ± nan	nan ± nan	0.0%	22/23
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	2.39904 ± 0.0339	816977.1000 ± 8385.0337	0.0%	0.0061 ± 0.0044	2002.8000 ± 1377.8644	0.4464 ± 0.2158	153400.0000 ± 75871.2067	100.0%	22/23
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	2.41486 ± 0.0513	816467.0000 ± 6914.5546	0.0%	0.0051 ± 0.0028	1678.7000 ± 853.0287	0.6450 ± 0.5411	219100.0000 ± 181662.5718	100.0%	22/22
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	2.40087 ± 0.0353	820875.1000 ± 4842.8374	0.0%	0.0042 ± 0.0023	1436.8000 ± 737.5390	1.1962 ± 0.6682	410222.2222 ± 229148.5582	90.0%	22/24
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	2.48184 ± 0.0243	840100.2000 ± 8147.4014	0.0%	0.0063 ± 0.0042	2036.6000 ± 1317.6898	nan ± nan	nan ± nan	0.0%	22/23
Q-learning ($\rho = 0.5, \epsilon = 1$)	10.01520 ± 0.1727	3441286.8000 ± 42249.9373	0.0%	0.0024 ± 0.0013	707.3000 ± 387.3637	0.7493 ± 0.4824	260100.0000 ± 167507.2834	100.0%	22/23
Q-learning ($\rho = 0.7, \epsilon = 1$)	10.01022 ± 0.0884	3453719.7000 ± 35047.7117	0.0%	0.0044 ± 0.0030	1478.9000 ± 956.0314	1.3257 ± 0.9365	464000.0000 ± 325482.7184	100.0%	22/23
Q-learning ($\rho = 0.9, \epsilon = 1$)	9.96116 ± 0.1227	3429200.0000 ± 43046.5147	0.0%	0.0042 ± 0.0037	1381.5000 ± 1143.9474	1.2725 ± 1.0050	437700.0000 ± 344933.9212	100.0%	22/23
Q-learning ($\rho = 0.99, \epsilon = 1$)	9.92413 ± 0.1156	3417322.0000 ± 31567.4616	0.0%	0.0041 ± 0.0047	957.2000 ± 1273.1382	nan ± nan	nan ± nan	0.0%	22/24
Async Value Iteration	0.11313 ± 0.0060	161800.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	22/23
Value Iteration	0.15098 ± 0.0044	213576.0000 ± 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 ρ 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 ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.29376 ± 0.0114	87332.2000 ± 130.8012	0.0%	0.0041 ± 0.0006	1229.7000 ± 55.2323	0.1200 ± 0.0058	35800.0000 ± 400.0000	100.0%	22/23
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.25831 ± 0.0102	80872.8000 ± 135.5838	0.0%	0.0040 ± 0.0007	1255.4000 ± 68.8421	0.0817 ± 0.0078	25600.0000 ± 489.8979	100.0%	22/22
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.25036 ± 0.0083	77700.9000 ± 434.0469	0.0%	0.0038 ± 0.0009	1217.4000 ± 47.3861	0.0654 ± 0.0032	20000.0000 ± 0.0000	100.0%	22/24
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.24895 ± 0.0086	76801.6000 ± 144.6556	0.0%	0.0036 ± 0.0009	1220.8000 ± 49.0649	0.0562 ± 0.0015	18000.0000 ± 0.0000	100.0%	22/22
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.35408 ± 0.0148	111584.0000 ± 412.9496	0.0%	0.0051 ± 0.0006	1864.1000 ± 289.8474	0.1230 ± 0.0039	39300.0000 ± 458.2576	100.0%	22/22
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.33255 ± 0.0036	105435.9000 ± 443.0337	0.0%	0.0046 ± 0.0009	1547.0000 ± 217.0396	0.1037 ± 0.0045	28700.0000 ± 458.2576	100.0%	22/22
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	0.32478 ± 0.0112	102077.8000 ± 445.8764	0.0%	0.0052 ± 0.0010	1663.0000 ± 226.2830	0.0724 ± 0.0037	23000.0000 ± 0.0000	100.0%	22/22
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	0.32277 ± 0.0091	102165.3000 ± 280.8591	0.0%	0.0051 ± 0.0013	1641.7000 ± 245.3341	0.0625 ± 0.0023	20300.0000 ± 458.2576	100.0%	22/22
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.51147 ± 0.0194	161853.1000 ± 806.6378	0.0%	0.0066 ± 0.0014	1961.0000 ± 305.9297	0.1420 ± 0.0061	46300.0000 ± 458.2576	100.0%	22/24
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.48351 ± 0.0082	156959.0000 ± 1504.6035	0.0%	0.0058 ± 0.0021	2019.1000 ± 561.3444	0.1063 ± 0.0051	34400.0000 ± 489.8979	100.0%	22/22
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	0.47548 ± 0.0124	154848.8000 ± 1231.8054	0.0%	0.0059 ± 0.0017	1840.0000 ± 576.0533	0.0844 ± 0.0030	28000.0000 ± 632.4555	100.0%	22/22
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.49243 ± 0.0149	158697.0000 ± 1032.6458	0.0%	0.0071 ± 0.0023	2187.3000 ± 527.6387	0.0792 ± 0.0069	25000.0000 ± 447.2136	100.0%	22/24
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	0.98622 ± 0.0094	328269.6000 ± 2378.9721	0.0%	0.0062 ± 0.0028	2125.7000 ± 859.0931	0.1737 ± 0.0042	59300.0000 ± 781.0250	100.0%	22/22
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	0.97079 ± 0.0225	322805.7000 ± 3815.3383	0.0%	0.0065 ± 0.0037	2080.7000 ± 1144.6155	0.1348 ± 0.0066	45800.0000 ± 600.0000	100.0%	22/22
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	1.00538 ± 0.0284	331516.4000 ± 2783.8548	0.0%	0.0070 ± 0.0021	2359.3000 ± 713.9961	0.1137 ± 0.0060	38100.0000 ± 700.0000	100.0%	22/22
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	1.03605 ± 0.0168	345653.4000 ± 2173.5444	0.0%	0.0075 ± 0.0020	2409.6000 ± 652.0138	0.1027 ± 0.0073	35000.0000 ± 1612.4515	100.0%	22/22
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	2.41348 ± 0.0200	823790.2000 ± 6985.4138	0.0%	0.0057 ± 0.0024	1960.4000 ± 799.8494	0.2159 ± 0.0126	74400.0000 ± 2576.8197	100.0%	22/22
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	2.32536 ± 0.1829	790112.1000 ± 61638.2405	30.0%	0.0064 ± 0.0038	2255.6000 ± 1445.1619	0.1717 ± 0.0096	59900.0000 ± 3112.8765	100.0%	22/22
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	2.41103 ± 0.3058	819447.7000 ± 104275.5428	40.0%	0.0061 ± 0.0035	2096.1000 ± 1082.7082	0.1480 ± 0.0119	50000.0000 ± 2366.4319	100.0%	22/22
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	2.33104 ± 0.3619	780753.1000 ± 126706.0884	50.0%	0.0059 ± 0.0040	1854.8000 ± 1323.8872	0.1392 ± 0.0100	47500.0000 ± 2459.6748	100.0%	22/24
Q-learning ($\rho = 0.5, \epsilon = 1$)	0.65477 ± 0.0912	227721.9000 ± 28488.3753	100.0%	0.0038 ± 0.0021	1269.4000 ± 920.4766	0.3416 ± 0.0252	118900.0000 ± 7777.5317	100.0%	22/22
Q-learning ($\rho = 0.7, \epsilon = 1$)	0.49519 ± 0.0859	172374.4000 ± 27628.8563	100.0%	0.0039 ± 0.0042	1199.7000 ± 992.5647	0.2737 ± 0.0198	96600.0000 ± 7989.9937	100.0%	22/22
Q-learning ($\rho = 0.9, \epsilon = 1$)	0.43197 ± 0.0747	153284.9000 ± 24918.4618	100.0%	0.0035 ± 0.0028	1076.7000 ± 854.1955	0.2381 ± 0.0333	85200.0000 ± 13362.6345	100.0%	22/22
Q-learning ($\rho = 0.99, \epsilon = 1$)	0.48867 ± 0.1156	168265.0000 ± 38578.7037	100.0%	0.0034 ± 0.0023	1149.3000 ± 862.5648	0.2304 ± 0.0338	81700.0000 ± 11081.9673	100.0%	22/22
Async Value Iteration	0.09015 ± 0.0031	132676.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	22/22
Value Iteration	0.12165 ± 0.0048	171508.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	22/22

Table 55. Comparing performance of Q-learning as we change the learning rate ρ in stochastic problem with $\gamma = 0.99$ (Problem 16). Two versions of value iteration are also appended for contrast

Algorithm (Problem 16) ($\gamma = 0.99$)	Runtime (mean ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.03071 ± 0.0094	11123.9000 ± 3373.6780	10.0%	0.0010 ± 0.0000	14.0000 ± 0.0000	0.0029 ± 0.0005	1100.0000 ± 300.0000	100.0%	5/5
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.03377 ± 0.0016	12223.1000 ± 13.9746	0.0%	0.0011 ± 0.0000	14.1000 ± 0.3000	0.0029 ± 0.0003	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.03690 ± 0.0029	12224.1000 ± 19.5880	0.0%	0.0009 ± 0.0000	13.7000 ± 0.6403	0.0030 ± 0.0006	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.03445 ± 0.0025	12243.3000 ± 32.2244	0.0%	nan ± nan	13.8000 ± 0.6000	0.0028 ± 0.0005	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.00332 ± 0.0010	1204.0000 ± 398.7759	100.0%	0.0010 ± 0.0000	17.7000 ± 3.4073	0.0033 ± 0.0010	1200.0000 ± 400.0000	100.0%	5/5
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.00366 ± 0.0014	1203.1000 ± 400.2021	100.0%	0.0010 ± 0.0000	15.1000 ± 2.3000	0.0035 ± 0.0011	1100.0000 ± 300.0000	100.0%	5/5
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	0.00275 ± 0.0005	1001.2000 ± 1.3266	100.0%	0.0010 ± 0.0000	16.7000 ± 4.5177	0.0028 ± 0.0005	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	0.00290 ± 0.0011	1101.3000 ± 299.5680	100.0%	nan ± nan	16.9000 ± 4.2059	0.0026 ± 0.0005	1000.0000 ± 0.0000	100.0%	5/6
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.00320 ± 0.0010	1102.7000 ± 299.1150	100.0%	nan ± nan	19.4000 ± 5.5172	0.0031 ± 0.0010	1100.0000 ± 300.0000	100.0%	5/5
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.00371 ± 0.0012	1204.7000 ± 399.1664	100.0%	0.0010 ± 0.0000	20.4000 ± 7.7872	0.0034 ± 0.0009	1100.0000 ± 300.0000	100.0%	5/5
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	0.00351 ± 0.0016	1204.8000 ± 598.7415	100.0%	nan ± nan	17.1000 ± 7.8543	0.0029 ± 0.0003	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.00355 ± 0.0010	1104.8000 ± 305.0832	100.0%	nan ± nan	16.3000 ± 6.0341	0.0032 ± 0.0009	1100.0000 ± 300.0000	100.0%	5/5
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	0.00317 ± 0.0012	1112.9000 ± 296.5186	100.0%	0.0011 ± 0.0000	21.5000 ± 11.8343	0.0032 ± 0.0012	1100.0000 ± 300.0000	100.0%	5/5
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	0.00343 ± 0.0011	1215.2000 ± 396.7883	100.0%	0.0010 ± 0.0000	27.7000 ± 18.3306	0.0031 ± 0.0007	1100.0000 ± 300.0000	100.0%	5/5
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	0.00305 ± 0.0011	1108.3000 ± 297.9806	100.0%	0.0010 ± 0.0000	16.8000 ± 6.0133	0.0027 ± 0.0004	1000.0000 ± 0.0000	100.0%	5/7
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	0.00284 ± 0.0004	1005.3000 ± 3.4943	100.0%	nan ± nan	20.9000 ± 11.4582	0.0028 ± 0.0004	1000.0000 ± 0.0000	100.0%	5/7
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	0.00424 ± 0.0022	1412.7000 ± 665.4364	100.0%	nan ± nan	19.8000 ± 13.0981	0.0035 ± 0.0014	1200.0000 ± 400.0000	100.0%	5/5
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	0.00400 ± 0.0017	1417.4000 ± 493.3733	100.0%	nan ± nan	19.5000 ± 12.4197	0.0031 ± 0.0010	1100.0000 ± 300.0000	100.0%	5/5
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	0.00335 ± 0.0012	1218.3000 ± 397.4521	100.0%	nan ± nan	24.0000 ± 24.6089	0.0026 ± 0.0005	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	0.00314 ± 0.0012	1107.5000 ± 297.6032	100.0%	0.0010 ± 0.0000	19.6000 ± 8.3929	0.0031 ± 0.0012	1100.0000 ± 300.0000	100.0%	5/7
Q-learning ($\rho = 0.5, \epsilon = 1$)	0.00322 ± 0.0010	1118.0000 ± 302.2552	100.0%	0.0010 ± 0.0000	20.5000 ± 10.8374	0.0029 ± 0.0003	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.7, \epsilon = 1$)	0.00289 ± 0.0015	1115.8000 ± 295.2446	100.0%	nan ± nan	21.6000 ± 12.5236	0.0026 ± 0.0008	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.9, \epsilon = 1$)	0.00322 ± 0.0010	1212.2000 ± 395.0410	100.0%	0.0010 ± 0.0000	28.6000 ± 27.8718	0.0030 ± 0.0010	1100.0000 ± 300.0000	100.0%	5/5
Q-learning ($\rho = 0.99, \epsilon = 1$)	0.00333 ± 0.0015	1226.1000 ± 393.2812	100.0%	nan ± nan	24.8000 ± 22.8070	0.0030 ± 0.0015	1100.0000 ± 300.0000	100.0%	5/5
Async Value Iteration	0.00040 ± 0.0005	200.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	5/5
Value Iteration	0.00070 ± 0.0005	320.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	5/5

Table 56. Comparing performance of Q-learning as we change the learning rate ρ 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 ± std)	#actions (mean ± std)	Convergence	Discover goal time (mean ± std)	Discover goal #actions (mean ± std)	Optimal Initial Cost2Go Time (mean ± std)	Optimal Initial Cost2Go #actions (mean ± std)	Optimal Initial Cost2Go Convergence	Shortest/Longest Path
Q-learning ($\rho = 0.5, \epsilon = 0$)	0.03311 ± 0.0011	12061.5000 ± 4.5880	0.0%	0.0010 ± 0.0000	14.0000 ± 0.0000	0.0025 ± 0.0007	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.7, \epsilon = 0$)	0.03330 ± 0.0014	12049.9000 ± 4.0608	0.0%	0.0010 ± 0.0000	14.0000 ± 0.0000	0.0026 ± 0.0007	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.9, \epsilon = 0$)	0.03523 ± 0.0021	12052.6000 ± 5.6604	0.0%	nan ± nan	14.0000 ± 0.0000	0.0032 ± 0.0010	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.99, \epsilon = 0$)	0.01518 ± 0.0180	4317.2000 ± 5066.6709	70.0%	nan ± nan	14.0000 ± 0.0000	0.0036 ± 0.0009	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.5, \epsilon = 0.25$)	0.00305 ± 0.0005	1001.7000 ± 1.3454	100.0%	0.0010 ± 0.0000	14.8000 ± 2.2271	0.0030 ± 0.0005	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.7, \epsilon = 0.25$)	0.00296 ± 0.0005	1002.9000 ± 1.6401	100.0%	0.0011 ± 0.0001	15.8000 ± 2.6000	0.0030 ± 0.0005	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.9, \epsilon = 0.25$)	0.00316 ± 0.0003	1002.9000 ± 2.5080	100.0%	nan ± nan	16.2000 ± 2.0881	0.0032 ± 0.0003	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.99, \epsilon = 0.25$)	0.00303 ± 0.0004	1002.0000 ± 1.4832	100.0%	nan ± nan	15.4000 ± 2.0100	0.0030 ± 0.0004	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.5, \epsilon = 0.5$)	0.00330 ± 0.0005	1005.7000 ± 3.3481	100.0%	nan ± nan	16.2000 ± 5.4000	0.0033 ± 0.0005	1000.0000 ± 0.0000	100.0%	5/7
Q-learning ($\rho = 0.7, \epsilon = 0.5$)	0.00352 ± 0.0012	1104.6000 ± 303.8109	100.0%	nan ± nan	20.2000 ± 5.7585	0.0032 ± 0.0009	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.9, \epsilon = 0.5$)	0.00316 ± 0.0006	1002.5000 ± 2.2023	100.0%	nan ± nan	16.0000 ± 3.7947	0.0032 ± 0.0006	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.99, \epsilon = 0.5$)	0.00303 ± 0.0011	1007.1000 ± 3.5903	100.0%	0.0011 ± 0.0000	21.4000 ± 5.4443	0.0030 ± 0.0011	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.5, \epsilon = 0.75$)	0.00305 ± 0.0001	1008.3000 ± 6.6340	100.0%	0.0010 ± 0.0000	25.0000 ± 13.7186	0.0031 ± 0.0001	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.7, \epsilon = 0.75$)	0.00367 ± 0.0014	1009.0000 ± 7.9246	100.0%	nan ± nan	20.4000 ± 11.8254	0.0036 ± 0.0014	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.9, \epsilon = 0.75$)	0.00370 ± 0.0008	1007.3000 ± 5.5687	100.0%	0.0010 ± 0.0000	15.6000 ± 8.2849	0.0036 ± 0.0008	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.99, \epsilon = 0.75$)	0.00354 ± 0.0012	1104.9000 ± 299.0797	100.0%	0.0010 ± 0.0000	21.2000 ± 9.6830	0.0031 ± 0.0003	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.5, \epsilon = 0.9$)	0.00332 ± 0.0005	1012.6000 ± 8.7430	100.0%	nan ± nan	18.7000 ± 9.1657	0.0033 ± 0.0005	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.7, \epsilon = 0.9$)	0.00271 ± 0.0004	1010.9000 ± 9.3429	100.0%	0.0010 ± 0.0000	23.6000 ± 10.6883	0.0026 ± 0.0004	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.9, \epsilon = 0.9$)	0.00337 ± 0.0007	1016.5000 ± 12.4197	100.0%	nan ± nan	20.0000 ± 7.3212	0.0034 ± 0.0007	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.99, \epsilon = 0.9$)	0.00326 ± 0.0007	1110.8000 ± 296.5181	100.0%	0.0010 ± 0.0000	16.6000 ± 6.6963	0.0030 ± 0.0005	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.5, \epsilon = 1$)	0.00303 ± 0.0007	1022.8000 ± 16.5638	100.0%	nan ± nan	19.4000 ± 22.7517	0.0029 ± 0.0005	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.7, \epsilon = 1$)	0.00290 ± 0.0003	1021.5000 ± 16.7108	100.0%	0.0010 ± 0.0000	35.3000 ± 37.5341	0.0028 ± 0.0004	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.9, \epsilon = 1$)	0.00271 ± 0.0005	1017.1000 ± 14.4599	100.0%	nan ± nan	18.4000 ± 7.3648	0.0026 ± 0.0005	1000.0000 ± 0.0000	100.0%	5/5
Q-learning ($\rho = 0.99, \epsilon = 1$)	0.00275 ± 0.0005	1024.8000 ± 29.6068	100.0%	0.0010 ± 0.0000	19.6000 ± 13.1697	0.0026 ± 0.0005	1000.0000 ± 0.0000	100.0%	5/5
Async Value Iteration	0.00030 ± 0.0005	144.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	5/5
Value Iteration	0.00061 ± 0.0005	224.0000 ± 0.0000	N/A	N/A	N/A	N/A	N/A	N/A	5/5