Appendix to *Lookahead-Bounded Q-Learning*

A Proofs

A.1 Proof of Proposition 2

Proof. We provide a proof that is similar to that of Proposition 2.3 (iv) of (Brown et al., 2010) but for the case of the absorption time formulation of an infinite horizon problem. Here, we define a policy $\pi := \{\pi_t\}_{t\geq 0}$ as a sequence of functions, that maps from $\{w_t\}_{t\geq 1}$ to feasible actions. We may also use stationary policies where π_t is the same for all t and only depends on the current state s_t. Let $\mathbb{G} = \{\mathcal{G}_t\}_{t\geq 0}$ be the perfect information relaxation of the natural filtration $\mathbb{F} = \{\mathcal{F}_t\}_{t\geq 0}$. Under $\mathbb{G}$, we have $\mathcal{G}_t = \mathcal{F}$, i.e., we have access to the entire future uncertainties at each t. Define by $\Pi_{\mathbb{G}}$ the set of policies that includes the policies that have access to future uncertainties in addition to nonanticipative policies. Let $\hat{\mathbb{G}}$ be a relaxation of $\mathbb{G}$ such that in addition to what is known under $\mathbb{G}$ the estimate penalty terms $\hat{\zeta}_t^{\pi_\varphi}(s_t, a_t, w_{t+1} \,|\, \varphi)$ are revealed at time t.

We first prove $\mathbf{E}[\hat{Q}_0^L(s,a)] \leq Q^*(s,a)$. For an admissible policy π, we have

$$
\begin{aligned}
\mathbf{E}[\hat{Q}_0^L(s,a)] &\overset{(a)}{=} \mathbf{E}\left[\sum_{t=0}^{\tau-1} r(s_t, \pi(s_t)) - \hat{\zeta}_t^{\pi}(s_t, a_t, w_{t+1} \,|\, \varphi) \,|\, s_0 = s,\, a_0 = a\right] \\
&\overset{(b)}{=} \mathbf{E}\left[\sum_{t=0}^{\tau-1} (r(s_t, \pi(s_t)) - \hat{\zeta}_t^{\pi}(s_t, a_t, w_{t+1} \,|\, \varphi))\, \mathbb{1}_{\{\tau<\infty\}} \,|\, s_0 = s,\, a_0 = a\right] \\
&\overset{(c)}{=} \sum_{\tau'=1}^{\infty} \mathbf{E}\left[\sum_{t=0}^{\tau'-1} (r(s_t, \pi(s_t)) - \hat{\zeta}_t^{\pi}(s_t, a_t, w_{t+1} \,|\, \varphi))\, \mathbb{1}_{\{\tau=\tau'\}} \,|\, s_0 = s,\, a_0 = a\right] \\
&\overset{(d)}{=} \sum_{\tau'=1}^{\infty} \mathbf{E}\left[\sum_{t=0}^{\tau'-1} (r(s_t, \pi(s_t)) - \mathbf{E}[\hat{\zeta}_t^{\pi}(s_t, a_t, w_{t+1} \,|\, \varphi) \,|\, \mathcal{G}_t])\, \mathbb{1}_{\{\tau=\tau'\}} \,|\, s_0 = s,\, a_0 = a\right] \\
&\overset{(e)}{=} \sum_{\tau'=1}^{\infty} \mathbf{E}\left[\sum_{t=0}^{\tau'-1} (r(s_t, \pi(s_t)) - \zeta_t^{\pi}(s_t, a_t, w_{t+1} \,|\, \varphi))\, \mathbb{1}_{\{\tau=\tau'\}} \,|\, s_0 = s,\, a_0 = a\right] \\
&\overset{(f)}{=} \mathbf{E}\left[\sum_{t=0}^{\tau-1} (r(s_t, \pi(s_t)) - \zeta_t^{\pi}(s_t, a_t, w_{t+1} \,|\, \varphi))\, \mathbb{1}_{\{\tau<\infty\}} \,|\, s_0 = s,\, a_0 = a\right] \\
&\overset{(g)}{=} \mathbf{E}\left[\sum_{t=0}^{\tau-1} (r(s_t, \pi(s_t)) - \zeta_t^{\pi}(s_t, a_t, w_{t+1} \,|\, \varphi)) \,|\, s_0 = s,\, a_0 = a\right] \\
&\leq Q^*(s,a)
\end{aligned}
$$

Equality (a) follows from the definition of $\hat{Q}_0(s,a)$ and equality (b) follows since τ has finite mean and r and φ are uniformly bounded. Equalities (c) and (d) follow from the law of total expectations. Equality (e) follows from Lemma A.1 in (Brown et al., 2010) and from the estimated penalty terms being unbiased, i.e., $\mathbf{E}[\hat{\zeta}_t^{\pi_G^*}(s_t, a_t, w_{t+1} \,|\, \varphi) \,|\, \mathcal{G}_t] = \zeta_t^{\pi_G^*}(s_t, a_t, w_{t+1} \,|\, \varphi)$. Equalities (f) and (g) follow by the law of total expectation and τ being almost surely finite stopping time, respectively. The inequality follows since the expected value of the penalty terms for a feasible policy is zero and the action-value function of a feasible policy, $Q^\pi(s,a)$, is less than $Q^*(s,a)$.

Now, we prove $Q^*(s,a) \leq \mathbf{E}[\hat{Q}_0^U(s,a)]$. Let π_G^* be the optimal solution for the dual problem,

$$
\max_{\pi_G \in \Pi_{\mathbb{G}}} \mathbf{E}\left[\sum_{t=0}^{\tau-1} (r(s_t, \pi_G) - \zeta_t^{\pi_G}(s_t, a_t, w_{t+1} \,|\, \varphi)) \,|\, s_0 = s,\, a_0 = a\right]. \tag{18}
$$

We have,

$$
\begin{aligned}
\mathbf{E}[\hat{Q}_0^U(s,a)] &\overset{(a)}{=} \mathbf{E}\left[\max_{\mathbf{a}}\left\{\sum_{t=0}^{\tau-1} r(s_t,a_t) - \hat{\zeta}_t^{\pi_\varphi}(s_t,a_t,w_{t+1}\,|\,\varphi)\right\} \,|\, s_0=s,\, a_0=a\right] \\
&\overset{(b)}{=} \max_{\pi_G\in\Pi_{\hat{\mathbb{G}}}} \mathbf{E}\left[\sum_{t=0}^{\tau-1}(r(s_t,\pi_G) - \hat{\zeta}_t^{\pi_G}(s_t,a_t,w_{t+1}\,|\,\varphi)) \,|\, s_0=s,\, a_0=a\right] \\
&\geq \mathbf{E}\left[\sum_{t=0}^{\tau-1}(r(s_t,\pi_G^*) - \hat{\zeta}_t^{\pi_G^*}(s_t,a_t,w_{t+1}\,|\,\varphi)) \,|\, s_0=s,\, a_0=a\right] \\
&\overset{(c)}{=} \mathbf{E}\left[\sum_{t=0}^{\tau-1}(r(s_t,\pi_G^*) - \hat{\zeta}_t^{\pi_G^*}(s_t,a_t,w_{t+1}\,|\,\varphi))\,\mathbb{1}_{\{\tau<\infty\}} \,|\, s_0=s,\, a_0=a\right] \\
&\overset{(d)}{=} \sum_{\tau'=1}^{\infty}\mathbf{E}\left[\sum_{t=0}^{\tau'-1}(r(s_t,\pi_G^*) - \hat{\zeta}_t^{\pi_G^*}(s_t,a_t,w_{t+1}\,|\,\varphi))\,\mathbb{1}_{\{\tau=\tau'\}} \,|\, s_0=s,\, a_0=a\right] \\
&\overset{(e)}{=} \sum_{\tau'=1}^{\infty}\mathbf{E}\left[\sum_{t=0}^{\tau'-1}(r(s_t,\pi_G^*) - \mathbf{E}[\hat{\zeta}_t^{\pi_G^*}(s_t,a_t,w_{t+1}\,|\,\varphi)\,|\,\mathcal{G}_t])\,\mathbb{1}_{\{\tau=\tau'\}} \,|\, s_0=s,\, a_0=a\right] \\
&\overset{(f)}{=} \sum_{\tau'=1}^{\infty}\mathbf{E}\left[\sum_{t=0}^{\tau'-1}(r(s_t,\pi_G^*) - \zeta_t^{\pi_G^*}(s_t,a_t,w_{t+1}\,|\,\varphi))\,\mathbb{1}_{\{\tau=\tau'\}} \,|\, s_0=s,\, a_0=a\right] \\
&\overset{(g)}{=} \mathbf{E}\left[\sum_{t=0}^{\tau-1}(r(s_t,\pi_G^*) - \zeta_t^{\pi_G^*}(s_t,a_t,w_{t+1}\,|\,\varphi))\,\mathbb{1}_{\{\tau<\infty\}} \,|\, s_0=s,\, a_0=a\right] \\
&\overset{(h)}{=} \mathbf{E}\left[\sum_{t=0}^{\tau-1}(r(s_t,\pi_G^*) - \zeta_t^{\pi_G^*}(s_t,a_t,w_{t+1}\,|\,\varphi)) \,|\, s_0=s,\, a_0=a\right] \\
&\overset{(i)}{=} \max_{\pi_G\in\Pi_{\mathbb{G}}} \mathbf{E}\left[\sum_{t=0}^{\tau-1}(r(s_t,\pi_G) - \zeta_t^{\pi_G}(s_t,a_t,w_{t+1}\,|\,\varphi)) \,|\, s_0=s,\, a_0=a\right] \\
&\geq Q^*(s,a).
\end{aligned}
$$

Equality (a) and (b) follow from the definition of $\hat{Q}_0^U(s,a)$ and since $\hat{\mathbb{G}}$ is a relaxation of the perfect information relaxation $\mathbb{G}$, which allows us to interchange the maximum and the expectation. The first inequality follows because $\pi_G^* \in \Pi_{\hat{\mathbb{G}}}$ since $\Pi_{\mathbb{G}} \subseteq \Pi_{\hat{\mathbb{G}}}$. Equality (c) follows since r and φ are uniformly bounded and τ has finite mean. Equalities (d) and (e) follow from the law of total expectations. Equality (f) follows from Lemma A.1 in (Brown et al., 2010) and from the estimated penalty terms being unbiased, i.e., $\mathbf{E}[\hat{\zeta}_t^{\pi_G^*}(s_t,a_t,w_{t+1}\,|\,\varphi)\,|\,\mathcal{G}_t] = \zeta_t^{\pi_G^*}(s_t,a_t,w_{t+1}\,|\,\varphi)$. Equalities (g) and (h) follow by the law of total expectation and τ being almost surely finite stopping time, respectively. Equality (i) follows since by definition π_G^* is the optimal solution of (18). The last inequality follows by weak duality (Proposition 1(i)). □

First, we state a technical lemma that is used in the proof of Proposition 3 and Lemma 1.

Lemma A.1. *For all $n = 1, 2, \ldots$, if $L_{n-1}(s,a) \leq U_{n-1}(s,a)$ and $Q'_{n-1} \in \mathcal{Q}$ then $L_n(s,a) \leq U_n(s,a)$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$.*

Proof. Fix an $(s,a) \in \mathcal{S}\times\mathcal{A}$. Note that the optimal values of the inner problems in (9) and (10), $\hat{Q}_0^U(s,a)$ and $\hat{Q}_0^L(s,a)$ respectively, are computed using the same sample path $\mathbf{w}$ and for each period within the inner DP, the same batch of samples is used for estimating the expectation in both the upper and lower bound problems. For clarity, let us denote the values of $\hat{Q}_0^L(s,a)$ and $\hat{Q}_0^U(s,a)$ at iteration $n = 1, 2, \ldots$ by $\hat{Q}_{n,0}^L(s,a)$ and $\hat{Q}_{n,0}^U(s,a)$, respectively. Assume $\alpha_n(s,a) \leq 1$ for all n. We provide a proof by induction. For $n = 1$, we have:

$$
Q_1(s_0,a_0) = Q'_0(s_0,a_0) + \alpha_0(s_0,a_0)\Big[r(s_0,a_0) + \gamma \max_a Q'_0(s_1,a) - Q'_0(s_0,a_0)\Big].
$$

Since the rewards $r(s,a)$ are uniformly bounded, $0<\gamma<1$ and $|Q'_0(s,a)|\leq\rho$ then Q_1 is bounded. Set $\varphi=Q_1$, since the actions selected by the policy π_{Q_1} are feasible in (9), we have

$$\hat{Q}^U_{1,0}(s,a)-\hat{Q}^L_{1,0}(s,a)\geq 0,$$

and with $L_0(s,a)\leq U_0(s,a)$, it follows that $L_1(s,a)\leq U_1(s,a)$. A similar proof can be used to show the inductive case also holds at iteration n,

$$\begin{aligned}Q_n(s_{n-1},a_{n-1}) &= Q'_{n-1}(s_{n-1},a_{n-1})\\ &+\alpha_{n-1}(s_{n-1},a_{n-1})\Big[r(s_{n-1},a_{n-1})+\gamma\max_a Q'_{n-1}(s_n,a)-Q'_{n-1}(s_{n-1},a_{n-1})\Big].\end{aligned}$$

By the inductive hypothesis, we have $Q'_{n-1}\in\mathcal{Q}$ and $L_{n-1}(s,a)\leq U_{n-1}(s,a)$. Then, similar to the base case, we have $Q_n\in\mathcal{Q}$ and $\hat{Q}^U_{n,0}(s,a)-\hat{Q}^L_{n,0}(s,a)\geq 0$. Therefore, $L_n(s,a)\leq U_n(s,a)$. Since our choice of (s,a) was arbitrary then the result follows for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. □

A.2 Proof of Proposition 3

Proof. Part (i): First, note that by (12) and (13) the upper and lower bound estimates $U_n(s,a)$ and $L_n(s,a)$ are bounded below and above by ρ and $-\rho$ respectively for all $(s,a)\in\mathcal{S}\times\mathcal{A}$ and for all n, where $\rho=R_{\max}/(1-\gamma)$. We assume in this proof that $\alpha_n(s,a)\leq 1$ for all n. Let $\tilde{L}_n=\max_{(s,a)}L_n(s,a)$ and $\tilde{U}_n=\max_{(s,a)}U_n(s,a)$. We claim that for every iteration n, we have that for all (s,a),

$$\bar{L}_n\leq Q_n(s,a)\leq\bar{U}_n\quad\text{and}\quad\bar{L}'_n\leq Q'_n(s,a)\leq\bar{U}'_n \tag{19}$$

where

$$\bar{L}_n=\min\left\{\tilde{U}_{n-1}(1+\gamma),\ldots,\tilde{U}_1\textstyle\sum_{i=0}^{n-1}\gamma^i,-M\sum_{i=0}^{n}\gamma^i\right\} \tag{20}$$

$$\bar{U}_n=\max\left\{\tilde{L}_{n-1}(1+\gamma),\ldots,\tilde{L}_1\textstyle\sum_{i=0}^{n-1}\gamma^i,M\sum_{i=0}^{n}\gamma^i\right\}, \tag{21}$$

$$\bar{L}'_n=\min\left\{\tilde{U}_n,\tilde{U}_{n-1}(1+\gamma),\ldots,\tilde{U}_1\textstyle\sum_{i=0}^{n-1}\gamma^i,-M\sum_{i=0}^{n}\gamma^i\right\}, \tag{22}$$

$$\bar{U}'_n=\max\left\{\tilde{L}_n,\tilde{L}_{n-1}(1+\gamma),\ldots,\tilde{L}_1\textstyle\sum_{i=0}^{n-1}\gamma^i,M\sum_{i=0}^{n}\gamma^i\right\}, \tag{23}$$

and M is a finite positive scalar defined as $M=\max\left\{R_{\max},\ \max_{(s,a)}Q_0(s,a)\right.$.

The result follows from the claim in (19). To see this note that at any iteration n, $\bar{L}_n$ and $\bar{L}'_n$ are bounded below by $-\rho\sum_{i=0}^{n}\gamma^i$ since each term inside the minimum of (20) and (22) is bounded below by $-\rho\sum_{i=0}^{n}\gamma^i$. As $n\to\infty$, we have

$$\frac{-\rho}{1-\gamma}\leq\liminf_{n\to\infty}\bar{L}_n\quad\text{and}\quad\frac{-\rho}{1-\gamma}\leq\liminf_{n\to\infty}\bar{L}'_n. \tag{24}$$

An analogous argument yields

$$\limsup_{n\to\infty}\bar{U}_n\leq\frac{\rho}{1-\gamma}\quad\text{and}\quad\limsup_{n\to\infty}\bar{U}'_n\leq\frac{\rho}{1-\gamma}. \tag{25}$$

Boundedness of $Q_n(s,a)$ and $Q'_n(s,a)$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$ follows from (19), (24) and (25).

Now, we prove our claim in (19) by induction. Since Algorithm 1 is asynchronous, at the nth iteration, the updates for the action-value iterates for (s,a), $Q_{n+1}(s,a)$ and $Q'_{n+1}(s,a)$, are either according to (11) and (14) (case 1) or set equal to $Q_n(s,a)$ and $Q'_n(s,a)$ respectively (case 2).

We first focus on $Q'_n(s,a)\leq\bar{U}_n$ part of (19), since $\bar{L}'_n\leq Q'_n(s,a)$ and $\bar{L}_n\leq Q_n(s,a)\leq\bar{U}_n$ proceed in an analogous manner. For $n=1$, we have $Q'_0(s,a)=Q_0(s,a)$, so if the update is carried out as in case 1,

$$\begin{aligned}Q_1(s,a) &= (1-\alpha_0(s,a))\,Q'_0(s,a)+\alpha_0(s,a)\,[r(s,a)+\gamma\max_a Q'_0(s',a)]\\ &\leq (1-\alpha_0(s,a))M+\alpha_0(s,a)M+\alpha_0(s,a)\gamma M\\ &\leq M(1+\gamma)\end{aligned}$$

so $Q'_1(s,a) \leq \max\{L_1(s,a), \min\{U_1(s,a), M(1+\gamma)\}\}$. Now consider the case where $U_1(s,a) \leq M(1+\gamma)$. Since Q_0 is bounded by ρ and $L_0(s,a) \leq U_0(s,a)$ then by Lemma A.1, we have $L_1(s,a) \leq U_1(s,a)$, so

$$Q'_1(s,a) \leq U_1(s,a) \leq M(1+\gamma). \tag{26}$$

Otherwise, if $U_1(s,a) \geq M(1+\gamma)$, we then have

$$Q'_1(s,a) \leq \max\{L_1(s,a), M(1+\gamma)\}. \tag{27}$$

From (26) and (27), we have

$$\begin{aligned} Q'_1(s,a) &\leq \max\{L_1(s,a), M(1+\gamma)\} \\ &\leq \max\{\tilde{L}_1, M(1+\gamma)\}. \end{aligned} \tag{28}$$

If the update is carried out as in case 2, we have,

$$\begin{aligned} Q'_1(s,a) &= Q'_0(s,a) \\ &\leq M \\ &< M(1+\gamma) \\ &\leq \max\{\tilde{L}_1, M(1+\gamma)\}. \end{aligned}$$

Thus $\bar{U}'_n(s,a)$ part of (19) is true for $n=1$. Suppose that it is true for $n = 1, 2, \ldots, k$. We will show it for $n = k+1$. Consider first the instance where the update is carried out according to case 1. We do casework on the inequality

$$Q'_k(s,a) \leq \max\left\{\tilde{L}_k, \tilde{L}_{k-1}(1+\gamma), \ldots, \tilde{L}_1 \textstyle\sum_{i=0}^{k-1} \gamma^i, M \sum_{i=0}^{k} \gamma^i\right\}, \tag{29}$$

which holds for all (s,a). First, let us consider the case where the right-hand-side of (29) is equal to $\tilde{L}_{k'} \sum_{i=0}^{k-k'} \gamma^i$ for some k' such that $1 \leq k' \leq k$. Then, we have

$$\begin{aligned} Q_{k+1}(s,a) &= (1-\alpha_k(s,a))Q'_k(s,a) + \alpha_k(s,a)[r(s,a) + \gamma \max_a Q'_k(s',a)] \\ &\leq (1-\alpha_k(s,a))\tilde{L}_{k'} \textstyle\sum_{i=0}^{k-k'} \gamma^i + \alpha_k(s,a)M + \alpha_k(s,a)\gamma\tilde{L}_{k'} \sum_{i=0}^{k-k'} \gamma^i \\ &\leq (1-\alpha_k(s,a))\tilde{L}_{k'} \textstyle\sum_{i=0}^{k-k'} \gamma^i + \alpha_k(s,a)\tilde{L}_{k'} + \alpha_k(s,a)\tilde{L}_{k'} \sum_{i=1}^{k-k'+1} \gamma^i \\ &= (1-\alpha_k(s,a))\tilde{L}_{k'} \textstyle\sum_{i=0}^{k-k'} \gamma^i + \alpha_k(s,a)\tilde{L}_{k'} \sum_{i=0}^{k-k'} \gamma^i + \alpha_k(s,a)\tilde{L}_{k'}\gamma^{k-k'+1} \\ &\leq \tilde{L}_{k'} \textstyle\sum_{i=0}^{k-k'} \gamma^i + \tilde{L}_{k'}\gamma^{k-k'+1} \\ &= \tilde{L}_{k'} \textstyle\sum_{i=0}^{k-k'+1} \gamma^i \end{aligned} \tag{30}$$

The first inequality holds by the induction assumption (29). The second inequality holds since in this case we have the right-hand-side of (29) is equal to $\tilde{L}_{k'}(1+\gamma+\ldots+\gamma^{k-k'})$. It follows that

$$\tilde{L}_{k'}(1+\gamma+\ldots+\gamma^{k-k'}) \geq M(1+\gamma+\ldots+\gamma^k),$$

which implies that $\tilde{L}_{k'} \geq M$. Finally, the third inequality holds by the assumption that $\alpha_n(s,a) \leq 1$ for all n. We have

$$\begin{aligned} Q'_{k+1}(s,a) &= \max\{L_{k+1}(s,a), \min\{U_{k+1}(s,a), Q_{k+1}(s,a)\}\} \\ &\leq \max\{L_{k+1}(s,a), \min\{U_{k+1}(s,a), \tilde{L}_{k'}(1+\gamma+\ldots+\gamma^{k-k'+1})\}\}. \end{aligned}$$

Now, consider the case where $U_{k+1}(s,a) \leq \tilde{L}_{k'}(1+\gamma+\ldots+\gamma^{k-k'+1})$. By the induction assumption, $Q'_n(s,a)$ is bounded below by $-\rho\sum_{i=0}^{n}\gamma^i$ and above by $\rho\sum_{i=0}^{n}\gamma^i$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and all $n = 1, 2, \ldots, k$. Since $L_0(s,a) \leq U_0(s,a)$, Lemma A.1 can be applied iteratively on $n = 1, \ldots, k+1$ to obtain that $L_{K+1}(s,a) \leq U_{k+1}(s,a)$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. Thus, we have

$$Q'_{k+1}(s,a) \leq U_{k+1}(s,a) \leq \tilde{L}_{k'}(1+\gamma+\ldots+\gamma^{k-k'+1}). \tag{31}$$

Otherwise, if $U_{k+1}(s,a) \geq \tilde{L}_{k'}(1+\gamma+\ldots+\gamma^{k-k'+1})$, we have

$$\begin{aligned} Q'_{k+1}(s,a) &\leq \max\{L_{k+1}(s,a), \tilde{L}_{k'}(1+\gamma+\ldots+\gamma^{k-k'+1})\} \\ &\leq \max\{\tilde{L}_{k+1}, \tilde{L}_{k'}(1+\gamma+\ldots+\gamma^{k-k'+1})\}. \end{aligned} \tag{32}$$

Moving on to the case where the right-hand-side of (29) is equal to $M(1+\gamma+\ldots+\gamma^k)$:

$$\begin{aligned} Q_{k+1}(s,a) &= (1-\alpha_k(s,a))\, Q'_k(s,a) + \alpha_k(s,a)\,[r(s,a) + \gamma \max_a Q'_k(s',a)] \\ &\leq (1-\alpha_k(s,a))M\textstyle\sum_{i=0}^{k}\gamma^i + \alpha_k(s,a)M + \alpha_k(s,a)\gamma M\sum_{i=0}^{k}\gamma^i \\ &= (1-\alpha_k(s,a))M\textstyle\sum_{i=0}^{k}\gamma^i + \alpha_k(s,a)M + \alpha_k(s,a)M\sum_{i=1}^{k+1}\gamma^i \\ &\leq M\textstyle\sum_{i=0}^{k}\gamma^i - \alpha_k(s,a)M\sum_{i=0}^{k}\gamma^i + \alpha_k(s,a)M\sum_{i=0}^{k}\gamma^i + M\gamma^{k+1} \\ &= M(1+\gamma+\ldots+\gamma^{k+1}). \end{aligned} \tag{33}$$

We have

$$\begin{aligned} Q'_{k+1}(s,a) &= \max\{L_{k+1}(s,a), \min(U_{k+1}(s,a), Q_{k+1}(s,a))\} \\ &\leq \max\{L_{k+1}(s,a), \min(U_{k+1}(s,a), M(1+\gamma+\ldots+\gamma^{k+1})\}\}. \end{aligned}$$

Now if $U_{k+1}(s,a) \leq M(1+\gamma+\ldots+\gamma^{k+1})$, then by applying Lemma A.1 as before,

$$Q'_{k+1}(s,a) \leq U_{k+1}(s,a) \leq M(1+\gamma+\ldots+\gamma^{k+1}). \tag{34}$$

Otherwise, if $U_{k+1}(s,a) \geq M(1+\gamma+\ldots+\gamma^{k+1})$, we have

$$\begin{aligned} Q'_{k+1}(s,a) &\leq \max\{L_{k+1}(s,a), M(1+\gamma+\ldots+\gamma^{k+1})\} \\ &\leq \max\{\tilde{L}_{k+1}, M(1+\gamma+\ldots+\gamma^{k+1})\}. \end{aligned} \tag{35}$$

Now, if the update is carried out according to case 2,

$$\begin{aligned} Q'_{k+1}(s,a) &= Q'_K(s,a) \\ &\leq \max\{\tilde{L}_k, \tilde{L}_{k-1}(1+\gamma), \ldots, \tilde{L}_1\textstyle\sum_{i=0}^{k-1}\gamma^i, M\sum_{i=0}^{k}\gamma^i\} \\ &\leq \max\{\tilde{L}_{k+1}, (1+\gamma)\tilde{L}_k, \ldots, \tilde{L}_1\textstyle\sum_{i=0}^{k}\gamma^i, M\sum_{i=0}^{k+1}\gamma^i\}. \end{aligned} \tag{36}$$

By (31), (32), (34), (35) and (36), we have $Q'_{k+1}(s,a) \leq \bar{U}'_{k+1}$. A similar argument can be made to show $\bar{L}_n \leq Q'_n(s,a)$ and $\bar{L}_n \leq Q_n(s,a) \leq \bar{U}_n$, which completes the inductive proof. □

Proof. Part (ii): Fix an $(s,a) \in \mathcal{S}\times\mathcal{A}$. By part (i) we have the action-value iterates Q_n and Q'_n are bounded for all n. We denote the "sampling noise" term using

$$\xi^L_n(s,a) = \hat{Q}^L_{n,0}(s,a) - \mathbf{E}[\hat{Q}^L_{n,0}(s,a)].$$

We also define an accumulated noise process started at iteration ν by $W^L_{\nu,\nu}(s,a) = 0$, and

$$W^L_{n+1,\nu}(s,a) = (1-\alpha_n(s,a))\, W^L_{n,\nu}(s,a) + \alpha_n(s,a)\,\xi^L_{n+1}(s,a) \quad \forall\, n \geq \nu,$$

which averages noise terms together across iterations. Note that τ is an almost surely finite stopping time, the rewards $r(s,a)$ are uniformly bounded, and Q_{n+1} is also bounded (by part (i)). Then, $\hat{Q}^L_{n,0}$ is bounded by some random variable and so is the conditional variance of $\xi^L_n(s,a)$. Hence, Corollary 4.1 in (Bertsekas & Tsitsiklis, 1996) applies and it follows that

$$\lim_{n\to\infty} W^L_{n,\nu}(s,a) = 0 \quad \forall\, \nu \geq 0.$$

Let $\tilde{\nu}$ be large enough so that $\alpha_n(s,a) \leq 1$ for all $n \geq \tilde{\nu}$. We also define

$$Y^L_{\tilde{\nu}}(s,a) = \rho,$$

$$Y^L_{n+1}(s,a) = (1-\alpha_n(s,a))\, Y^L_n(s,a) + \alpha_n(s,a)\, Q^*(s,a), \quad \forall\, n \geq \tilde{\nu}.$$

It is easy to see that the sequence $Y^L_n(s,a) \to Q^*(s,a)$. We claim that for all iterations $n \geq \tilde{\nu}$, it holds that

$$L_n(s,a) \leq \min\{\rho,\, Y^L_n(s,a) + W^L_{n,\tilde{\nu}}(s,a)\}.$$

To prove this claim, we proceed by induction on n. For $n = \tilde{\nu}$, we have

$$Y^L_{\tilde{\nu}}(s,a) = \rho \quad \text{and} \quad W^L_{\tilde{\nu},\tilde{\nu}}(s,a) = 0,$$

so it is clear that the statement is true for the base case. We now show that it is true for $n+1$ given that it holds at n:

$$\begin{aligned}
L_{n+1}(s,a) &= \min\{\rho,\, (1-\alpha_n(s,a))\, L_n(s,a) + \alpha_n(s,a)\, (\hat{Q}^L_{n,0}(s,a) - \mathbf{E}[\hat{Q}^L_{n,0}(s,a)] + \mathbf{E}[\hat{Q}^L_{n,0}(s,a)])\} \\
&= \min\{\rho,\, (1-\alpha_n(s,a))\, L_n(s,a) + \alpha_n(s,a)\, \xi^L_n(s,a) + \alpha_n(s,a)\, \mathbf{E}[\hat{Q}^L_{n,0}(s,a)]\} \\
&\leq \min\{\rho,\, (1-\alpha_n(s,a))\, (Y^L_n(s,a) + W^L_{n,\nu_k}(s,a)) + \alpha_n(s,a)\, \xi^L_n(s,a) + \alpha_n(s,a)\, Q^*(s,a)\} \\
&\leq \min\{\rho,\, Y^L_{n+1}(s,a) + W^L_{n+1,\tilde{\nu}}(s,a)\},
\end{aligned}$$

where the first inequality follows by the induction hypothesis and $\mathbf{E}[\hat{Q}^L_{n,0}(s,a)] \leq Q^*(s,a)$ follows by Proposition 2. Next, since $Y^L_n(s,a) \to Q^*(s,a)$, $W^L_{n,\nu_k}(s,a) \to 0$ and $Q^*(s,a) \leq \rho$, we have

$$\limsup_{n\to\infty}\, L_n(s,a) \leq Q^*(s,a).$$

Therefore, since our choice of (s,a) was arbitrary, it follows that for every $\eta > 0$, there exists some time n' such that $L_n(s,a) \leq Q^*(s,a) + \eta$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $n \geq n'$.

Using Proposition 2, $Q^*(s,a) \leq \mathbf{E}[\hat{Q}^U_{n,0}(s,a)]$, a similar argument as the above can be used to establish that

$$Q^*(s,a) \leq \liminf_{n\to\infty}\, U_n(s,a).$$

Hence, there exists some time n'' such that $Q^*(s,a) - \eta \leq U_n(s,a)$ for all (s,a) and $n \geq n''$. Take n_0 to be some time greater than n' and n'' and the result follows. □

A.3 Proof of Lemma 1

Proof. We use induction on n. Since for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, $L_0(s,a) \leq U_0(s,a)$ and $-\rho \leq Q'_0(s,a) \leq \rho$ then $L_1(s,a) \leq U_1(s,a)$ by Lemma A.1. Suppose that $L_n(s,a) \leq U_n(s,a)$ holds for all (s,a) for all $n = 1, \ldots, k$. For all $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have $Q'_k(s,a)$ is bounded since by Proposition 3(i) $Q'_n(s,a)$ is bounded for all n. We also have $L_k(s,a) \leq U_k(s,a)$ for all (s,a) by the induction assumption. Applying Lemma A.1 again at $n = k+1$ yields $L_{k+1}(s,a) \leq U_{k+1}(s,a)$ and the inductive proof is complete. □

A.4 Proof of Theorem 1

Proof. We first prove part (i). We start by writing Algorithm 1 using DP operator notation. Define a mapping H such that

$$(HQ')(s,a) = r(s,a) + \gamma \mathbf{E}\left[\max_{a'} Q'(s',a')\right],$$

where $s' = f(s,a,w)$. It is well-known that the mapping H is a γ-contraction in the maximum norm. We also define a random noise term

$$\xi_n(s,a) = \gamma \max_{a'} Q'_n(s',a') - \gamma \mathbf{E}\left[\max_{a'} Q'_n(s',a')\right]. \tag{37}$$

The main update rules of Algorithm 1 can then be written as

$$\begin{aligned}
Q_{n+1}(s,a) &= (1-\alpha_n(s,a))\, Q'_n(s,a) + \alpha_n(s,a)\, \left[(HQ'_n)(s,a) + \xi_{n+1}(s,a)\right], \\
U_{n+1}(s,a) &= \Pi_{[-\rho,\,\infty]} \left[(1-\beta_n(s,a))\, U_n(s,a) + \beta_n(s,a)\, \hat{Q}^U_0(s,a)\right], \\
L_{n+1}(s,a) &= \Pi_{[\infty,\,\rho]} \left[(1-\beta_n(s,a))\, L_n(s,a) + \beta_n(s,a)\, \hat{Q}^L_0(s,a)\}\right],
\end{aligned}$$

$$Q'_{n+1}(s,a) = \Pi_{[L_{n+1}(s,a),\, U_{n+1}(s,a)]}\left[Q_{n+1}(s,a)\right]. \tag{38}$$

Assume without loss of generality that $Q^*(s,a) = 0$ for all state-action pairs (s,a). This can be established by shifting the origin of the coordinate system. Note that by (38) at any iteration n and for all (s,a), we have $L_n(s,a) \leq Q'_n(s,a) \leq U_n(s,a)$.

We proceed via induction. First, note that by Propostion 3(i) the iterates of Algorithm 1 $Q'_n(s,a)$ are bounded in the sense that there exists a constant D_0 such that $|Q'_n(s,a)| \leq D_0$ for all (s,a) and iterations n. Define the sequence $D_{k+1} = (\gamma + \epsilon)\, D_k$, such that $\gamma + \epsilon < 1$ and $\epsilon > 0$. Clearly, $D_k \to 0$. Suppose that there exists some time n_k such that for all (s,a),

$$\max\{-D_k, L_n(s,a)\} \leq Q'_n(s,a) \leq \min\{D_k, U_n(s,a)\}, \;\; \forall n \geq n_k.$$

We will show that this implies the existence of some time n_{k+1} such that

$$\max\{-D_{k+1}, L_n(s,a)\} \leq Q'_n(s,a) \leq \min\{D_{k+1}, U_n(s,a)\} \;\; \forall\, (s,a),\; n \geq n_{k+1}.$$

This implies that $Q'_n(s,a)$ converges to $Q^*(s,a) = 0$ for all (s,a). We also assume that $\alpha_n(s,a) \leq 1$ for all (s,a) and n. Define an accumulated noise process started at n_k by $W_{n_k,n_k}(s,a) = 0$, and

$$W_{n+1,n_k}(s,a) = (1 - \alpha_n(s,a))\, W_{n,n_k}(s,a) + \alpha_n(s,a)\, \xi_{n+1}(s,a), \quad \forall\, n \geq n_k, \tag{39}$$

where $\xi_n(s,a)$ is as defined in (37). We now use Corollary 4.1 in (Bertsekas & Tsitsiklis, 1996) which states that under the assumptions of Theorem 1 on the step size $\alpha_n(s,a)$, and if $\mathbf{E}[\xi_n(s,a) \,|\, \mathcal{F}_n] = 0$ and $\mathbf{E}[\xi_n^2(s,a) \,|\, \mathcal{F}_n] \leq A_n$, where the random variable A_n is bounded with probability 1, the sequence $W_{n+1,n_k}(s,a)$ defined in (39) converges to zero, with probability 1. From our definition of the stochastic approximation noise $\xi_n(s,a)$ in (37), we have

$$\mathbf{E}[\xi_n(s,a) \,|\, \mathcal{F}_n] = 0 \quad \text{and} \quad \mathbf{E}[\xi_n^2(s,a) \,|\, \mathcal{F}_n] \leq C(1 + \max_{s',a'} Q'^2_n(s',a')),$$

where C is a constant. Then, it follows that

$$\lim_{n\to\infty} W_{n,n_k}(s,a) = 0 \quad \forall\, (s,a),\; n_k.$$

Now, for the sake of completeness, we restate a lemma from (Bertsekas & Tsitsiklis, 1996) below, which we will use to bound the accumulated noise.

Lemma A.2 (Lemma 4.2 in (Bertsekas & Tsitsiklis, 1996))**.** *For every $\delta > 0$, and with probability one, there exists some n' such that $|W_{n,n'}(s,a)| \leq \delta$, for all $n \geq n'$.*

Using the above lemma, let $n_{k'} \geq n_k$ such that, for all $n \geq n_{k'}$ we have

$$|W_{n,n_{k'}}(s,a)| \leq \gamma\epsilon D_k < \gamma D_k.$$

Furthermore, by Proposition 3(ii) let $\nu_k \geq n_{k'}$ such that, for all $n \geq \nu_k$ we have

$$L_n(s,a) \leq \gamma D_k - \gamma\epsilon D_k \quad \text{and} \quad \gamma\epsilon D_k - \gamma D_k \leq U_n(s,a).$$

Define another sequence Y_n that starts at iteration ν_k.

$$Y_{\nu_k}(s,a) = D_k \quad \text{and} \quad Y_{n+1}(s,a) = (1 - \alpha_n(s,a))\, Y_n(s,a) + \alpha_n(s,a)\, \gamma\, D_k \tag{40}$$

Note that it is easy to show that the sequence $Y_n(s,a)$ in (40) is decreasing, bounded below by γD_k, and converges to γD_k as $n \to \infty$. Now we state the following lemma.

Lemma A.3. *For all state-action pairs (s,a) and iterations $n \geq \nu_k$, it holds that:*

(1) $Q'_n(s,a) \leq \min\{U_n(s,a), Y_n(s,a) + W_{n,\nu_k}(s,a)\}$,

(2) $\max\{L_n(s,a), -Y_n(s,a) + W_{n,\nu_k}(s,a)\} \leq Q'_n(s,a)$.

Proof. We focus on part (1). For the base case $n = \nu_k$, the statement holds because $Y_{\nu_k}(s,a) = D_k$ and $W_{\nu_k,\nu_k}(s,a) = 0$. We assume it is true for n and show that it continues to hold for $n+1$:

$$\begin{aligned}
Q_{n+1}(s,a) &= (1-\alpha_n(s,a))Q'_n(s,a) + \alpha_n(s,a)\,[(HQ'_n)(s,a) + \xi_{n+1}(s,a)] \\
&\le (1-\alpha_n(s,a))\min\{U_n(s,a), Y_n(s,a) + W_{n,\nu_k}(s,a)\} \\
&\qquad + \alpha_n(s,a)\,(HQ'_n)(s,a) + \alpha_n(s,a)\,\xi_{n+1}(s,a) \\
&\le (1-\alpha_n(s,a))\,(Y_n(s,a) + W_{n,\nu_k}(s,a)) + \alpha_n(s,a)\,\gamma D_k + \alpha_n(s,a)\,\xi_{n+1}(s,a) \\
&\le Y_{n+1}(s,a) + W_{n+1,\nu_k}(s,a),
\end{aligned}$$

where we used $(HQ'_n) \le \gamma\|Q'_n\| \le \gamma D_k$. Now, we have

$$\begin{aligned}
Q'_{n+1}(s,a) &= \Pi_{[L_{n+1}(s,a),\,U_{n+1}(s,a)]}\,[Q_{n+1}(s,a)] \\
&\le \Pi_{[L_{n+1}(s,a),\,U_{n+1}(s,a)]}\,[Y_{n+1}(s,a) + W_{n+1,\nu_k}(s,a)] \\
&\le \min\{U_{n+1}(s,a), Y_{n+1}(s,a) + W_{n+1,\nu_k}(s,a)\}.
\end{aligned}$$

The first inequality holds because

$$Q_{n+1}(s,a) \le Y_{n+1}(s,a) + W_{n+1,\nu_k}(s,a).$$

The second inequality holds because $Y_{n+1}(s,a) + W_{n+1,\nu_k}(s,a) \ge \gamma D_k - \gamma\epsilon D_k$, $L_n(s,a) \le \gamma D_k - \gamma\epsilon D_k$, and $L_n(s,a) \le U_n(s,a)$ by Lemma 1. Symmetrically, it can be shown that

$$\max\{L_{n+1}(s,a), -Y_{n+1}(s,a) + W_{n+1,\nu_k}(s,a)\} \le Q'_{n+1}(s,a),$$

which completes the proof. □

Since $Y_n(s,a) \to \gamma D_k$ and $W_{n,\nu_k}(s,a) \to 0$, we have

$$\limsup_{n\to\infty}\|Q'_n\| \le \gamma D_k < D_{k+1}.$$

Therefore, there exists some time n_{k+1} such that

$$\max\{-D_{k+1}, L_n(s,a)\} \le Q'_n(s,a) \le \min\{D_{k+1}, U_n(s,a)\}\ \ \forall\,(s,a),\ n \ge n_{k+1},$$

completing thus the induction.

For part (ii) of the theorem: we fix (s,a) and focus on the convergence analysis of $U_n(s,a)$ to $Q^*(s,a)$. A similar analysis can be done to show $L_n(s,a) \to Q^*(s,a)$ almost surely. First note that we can write (12) the update equation of $U_n(s,a)$ as:

$$U_{n+1}(s_n,a_n) = \Pi_{[-\rho,\,\infty]}\left[U_n(s_n,a_n) + \beta_n(s_n,a_n)\left[\psi_n(U_n(s_n,a_n), Q_{n+1}(s_n,a_n))\right]\right]$$

where $\psi_n(U_n(s_n,a_n), Q_{n+1}(s_n,a_n))$ is the stochastic gradient and in this case is equal to $\hat{Q}_0^U(s_n,a_n) - U_n(s_n,a_n)$. We define the noise terms

$$\bar{\epsilon}_{n+1}(s_n,a_n) = \psi_n(U_n(s_n,a_n), Q^*(s_n,a_n)) - \mathbf{E}[\psi_n(U_n(s_n,a_n), Q^*(s_n,a_n))] \tag{41}$$

$$\bar{\varepsilon}_{n+1}(s_n,a_n) = \psi_n(U_n(s_n,a_n), Q_{n+1}(s_n,a_n)) - \psi_n(U_n(s_n,a_n), Q^*(s_n,a_n)). \tag{42}$$

Note here that $\bar{\epsilon}_{n+1}(s_n,a_n)$ represents the error that the sample gradient deviates from its mean when computed using the optimal action-value Q^* and $\bar{\varepsilon}_{n+1}(s_n,a_n)$ is the error between the two evaluations of ψ_n due only to the difference between $Q_{n+1}(s_n,a_n)$ and $Q^*(s_n,a_n)$. Thus, we have

$$\psi_n(U_n(s_n,a_n), Q_{n+1}(s_n,a_n)) = \mathbf{E}[\psi_n(U_n(s_n,a_n), Q^*(s_n,a_n))] + \bar{\epsilon}_{n+1}(s_n,a_n) + \bar{\varepsilon}_{n+1}(s_n,a_n).$$

Since $Q_n \to Q^*$ by part (i) of the Theorem, then $\bar{\varepsilon}_n(s_n,a_n) \to 0$ almost surely. It is now convenient to view $U_n(s,a)$ as a stochastic process in n, adapted to the filtration $\{\mathcal{F}_n\}_{n\ge 0}$. By definition of $\bar{\epsilon}_{n+1}(s,a)$, we have that

$$\mathbf{E}[\bar{\epsilon}_{n+1}(s,a)|\mathcal{F}_n] = 0 \quad a.s.$$

Since $\bar{\epsilon}_{n+1}(s,a)$ is unbiased and $\bar{\varepsilon}_{n+1}(s,a)$ converges to zero, we can apply Theorem 2.4 of (Kushner & Yin, 2003), a standard stochastic approximation convergence result, to conclude that $U_n(s,a) \to Q^*(s,a)$ almost surely. Since our choice of (s,a) was arbitrary, this convergence holds for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. □

A.5 Proof of Lemma 2

Proof. First note that Lemma A.1 still holds in this case. To see this, note that if the requirements of the lemma are satisfied (i.e., if we are at iteration $n = 1, 2, \ldots$, and in the previous iteration, we had $L_{n-1}(s,a) \leq U_{n-1}(s,a)$ for all (s,a) and $Q'_{n-1} \in \mathcal{Q}$), then Q_n is bounded using the same argument as before. Since the rewards $r(s,a)$ are uniformly bounded and τ is an almost surely finite stopping time, then $\tilde{Q}^L_{n,0}$ and $\tilde{Q}^U_{n,0}$ are finite. Moreover, since $\tilde{Q}^L_{n,0}$ and $\tilde{Q}^U_{n,0}$ are computed using the same sample path $\mathbf{w}$, it follows that

$$\tilde{Q}^U_{n,0}(s,a) - \tilde{Q}^L_{n,0}(s,a) \geq 0, \quad \text{for all } (s,a) \in \mathcal{S} \times \mathcal{A}.$$

This can be easily seen if we subtract (17) from (16). Notice that the reward and the penalty will both cancel out and we have $\tilde{Q}^U_{n,t} - \tilde{Q}^L_{n,t} \geq 0$ for all $t = 0, 1, \ldots, \tau - 1$. With $L_{n-1}(s,a) \leq U_{n-1}(s,a)$ and $\tilde{Q}^U_{n,0}(s,a) \geq \tilde{Q}^L_{n,0}(s,a)$ it follows that $L_n(s,a) \leq U_n(s,a)$ for all (s,a).

Now, we prove Proposition 3 again for the experience replay buffer case.

For part (i): our original proof still holds since Lemma A.1 still holds.

For part (ii): first note that since the experience replay buffer is updated with a new observation of the noise at every iteration, then by Borel's law of large numbers, we have our probability estimate $\hat{p}(w)$ for the noise converges to the true noise distribution $p(w)$ as $n \to \infty$, i.e.,

$$\lim_{n\to\infty} \hat{p}_n(w) = p(w) \text{ for all } w \in \mathcal{W}. \tag{43}$$

Fix an $(s,a) \in \mathcal{S} \times \mathcal{A}$. By part (i) we have the action-value iterates Q_n and Q'_n are bounded for all n. Now we write the iterate $\tilde{Q}^L_{n,0}(s,a)$ in terms of a noise term and a bias term as follows,

$$\tilde{Q}^L_{n,0}(s,a) = \underbrace{\tilde{Q}^L_{n,0}(s,a) - \mathbf{E}[\tilde{Q}^L_{n,0}(s,a)]}_{\text{noise}} + \underbrace{\mathbf{E}[\tilde{Q}^L_{n,0}(s,a)] - \mathbf{E}[Q^L_{n,0}(s,a)]}_{\text{bias}} + \mathbf{E}[Q^L_{n,0}(s,a)].$$

Now, we define the noise term using

$$\xi^L_n(s,a) = \tilde{Q}^L_{n,0}(s,a) - \mathbf{E}[\tilde{Q}^L_{n,0}(s,a)].$$

Also, similar to the original proof we define an accumulated noise process started at iteration ν by $W^L_{\nu,\nu}(s,a) = 0$, and

$$W^L_{n+1,\nu}(s,a) = (1 - \alpha_n(s,a))\, W^L_{n,\nu}(s,a) + \alpha_n(s,a)\, \xi^L_{n+1}(s,a) \quad \forall\, n \geq \nu,$$

which averages noise terms together across iterations. We have $\mathbf{E}[\tilde{Q}^L_{n,0}(s,a) - \mathbf{E}[\tilde{Q}^L_{n,0}(s,a)]|\mathcal{F}_n] = 0$, so Corollary 4.1 applies and it follows that

$$\lim_{n\to\infty} W^L_{n,\nu}(s,a) = 0 \quad \forall\, \nu \geq 0.$$

Let $\tilde{\nu}$ be large enough so that $\alpha_n(s,a) \leq 1$ for all $n \geq \tilde{\nu}$. We denote the bias term by

$$\chi_n(s,a) = \mathbf{E}[\tilde{Q}^L_{n,0}(s,a)] - \mathbf{E}[Q^L_{n,0}(s,a)].$$

Since as $n \to \infty$, we have $\hat{p}_n(w) \to p(w)$, the bias due to sampling from the experience buffer $\chi_n(s,a) \to 0$. Let $\eta > 0$ and $\bar{\nu} \geq \tilde{\nu}$ be such that $|\chi(s,a)| \leq \frac{\eta}{2}$ for all $n \geq \bar{\nu}$ and all (s,a). We also define

$$\begin{aligned} Y^L_{\tilde{\nu}}(s,a) &= \rho, \\ Y^L_{n+1}(s,a) &= (1 - \alpha_n(s,a))\, Y^L_n(s,a) + \alpha_n(s,a)\, Q^*(s,a) + \alpha_n(s,a)\, \frac{\eta}{2}, \quad \forall\, n \geq \bar{\nu}. \end{aligned}$$

It is easy to see that the sequence $Y^L_n(s,a) \to Q^*(s,a) + \frac{\eta}{2}$. Now we show that the following claim holds. Claim: for all iterations $n \geq \bar{\nu}$, it holds that

$$L_n(s,a) \leq \min\{\rho,\, Y^L_n(s,a) + W^L_{n,\bar{\nu}}(s,a)\}.$$

To prove this claim, we proceed by induction on n. For $n = \bar{\nu}$, we have

$$Y^L_{\bar{\nu}}(s,a) = \rho \quad \text{and} \quad W^L_{\bar{\nu},\bar{\nu}}(s,a) = 0,$$

so the statement is true for the base case. We now show that it is true for $n+1$ given that it holds at n:

$$\begin{aligned}
L_{n+1}(s,a) &= \min\{\rho,\ (1-\alpha_n(s,a))\, L_n(s,a) + \alpha_n(s,a)\,(\tilde{Q}^L_{n,0}(s,a) - \mathbf{E}[\tilde{Q}^L_{n,0}(s,a)] \\
&\qquad + \mathbf{E}[\tilde{Q}^L_{n,0}(s,a)] - \mathbf{E}[Q^L_{n,0}(s,a)] + \mathbf{E}[Q^L_{n,0}(s,a)])\} \\
&= \min\{\rho,\ (1-\alpha_n(s,a))\, L_n(s,a) + \alpha_n(s,a)\,\xi^L_n(s,a) + \alpha_n(s,a)\,\chi_n(s,a) \\
&\qquad + \alpha_n(s,a)\,\mathbf{E}[Q^L_{n,0}(s,a)]\} \\
&\le \min\{\rho,\ (1-\alpha_n(s,a))\,(Y^L_n(s,a) + W^L_{n,\nu_k}(s,a)) + \alpha_n(s,a)\,\xi^L_n(s,a) \\
&\qquad + \alpha_n(s,a)\,\frac{\eta}{2} + \alpha_n(s,a)\,Q^*(s,a)\} \\
&\le \min\{\rho,\ Y^L_{n+1}(s,a) + W^L_{n+1,\tilde{\nu}}(s,a)\},
\end{aligned}$$

where the first inequality follows by the induction hypothesis and $\mathbf{E}[Q^L_{n,0}(s,a)] \le Q^*(s,a)$. Next, since $Y^L_n(s,a) \to Q^*(s,a) + \frac{\eta}{2}$, $W^L_{n,\nu_k}(s,a) \to 0$, then if $Q^*(s,a) + \frac{\eta}{2} \le \rho$, we have

$$\limsup_{n\to\infty}\ L_n(s,a) \le Q^*(s,a) + \frac{\eta}{2}.$$

Otherwise, if $\rho < Q^*(s,a) + \frac{\eta}{2}$, then

$$\limsup_{n\to\infty}\ L_n(s,a) \le \rho < Q^*(s,a) + \frac{\eta}{2}.$$

Therefore, since our choice of (s,a) was arbitrary, it follows that for every $\eta > 0$, there exists some time n' such that $L_n(s,a) \le Q^*(s,a) + \eta$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $n \ge n'$.

Using Proposition 1(i), $Q^*(s,a) \le \mathbf{E}[Q^U_{n,0}(s,a)]$, a similar argument as the above can be used to establish that

$$Q^*(s,a) - \frac{\eta}{2} \le \liminf_{n\to\infty}\ U_n(s,a).$$

Hence, there exists some time n'' such that $Q^*(s,a) - \eta \le U_n(s,a)$ for all (s,a) and $n \ge n''$. Take n_0 to be some time greater than n' and n'' and the result follows.

The proof of Lemma 1 when using an experience buffer is similar to that given in appendix A.3 so it is omitted. □

A.6 Proof of Theorem 2

Proof. The proof of both parts (i) and (ii) are similar to that of Theorem 1, so we omit them. □

B LBQL with Experience Replay Algorithm

Algorithm 2 LBQL with Experience Replay

Input: Initial estimates $L_0 \le Q_0 \le U_0$, batch size K, stepsize rules $\alpha_n(s,a)$, $\beta_n(s,a)$, and noise buffer $\mathcal{B}$.
Output: Approximations $\{L_n\}$, $\{Q'_n\}$, and $\{U_n\}$.
Set $Q'_0 = Q_0$ and choose an initial state s_0.
for $n = 0, 1, 2, \ldots$ **do**
Choose an action a_n via some behavior policy (e.g., ϵ-greedy). Observe w_{n+1}.
Store w_{n+1} in $\mathcal{B}$ and update $\hat{p}_n(w_{n+1})$.
Perform a standard Q-learning update:

$$Q_{n+1}(s_n,a_n) = Q'_n(s_n,a_n) + \alpha_n(s_n,a_n)\Big[r_n(s_n,a_n) + \gamma \max_a Q'_n(s_{n+1},a) - Q'_n(s_n,a_n)\Big].$$

Sample randomly a sample path $\mathbf{w} = (w_1, w_2, \ldots, w_\tau)$ from $\mathcal{B}$, where $\tau \sim \mathrm{Geom}(1-\gamma)$.
Set $\varphi = Q_{n+1}$. Using $\mathbf{w}$ and the current $\hat{p}_n$ compute $\tilde{Q}_0^U(s_n,a_n)$ and $\tilde{Q}_0^L(s_n,a_n)$, using (16) and (17), respectively.
Update and enforce upper and lower bounds:

$$U_{n+1}(s_n,a_n) = \Pi_{[-\rho,\,\infty]}\Big[U_n(s_n,a_n) + \beta_n(s_n,a_n)\left[\tilde{Q}_0^U(s_n,a_n) - U_n(s_n,a_n)\right]\Big],$$
$$L_{n+1}(s_n,a_n) = \Pi_{[\infty,\,\rho]}\Big[L_n(s_n,a_n) + \beta_n(s_n,a_n)\left[\tilde{Q}_0^L(s_n,a_n) - L_n(s_n,a_n)\right]\Big],$$

$$Q'_{n+1}(s_n,a_n) = \Pi_{[L_{n+1}(s_n,a_n),\,U_{n+1}(s_n,a_n)]}\left[Q_{n+1}(s_n,a_n)\right]$$

end for

C Implementation Details of LBQL with Experience Replay

We use a noise buffer $\mathcal{B}$ of size κ to record the noise values w that have been previously observed. The buffer $\mathcal{B}$ is used to generate the sample path $\mathbf{w}$ and the batch sample $\{w_1, \ldots, w_K\}$ used to estimate the expectation in the penalty function. Here, it is not necessary that the noise space $\mathcal{W}$ is finite. This is also convenient in problems with a large noise support such as the car-sharing problem with four stations where we have two sources of noise. Specifically, the noise due to the distribution of the rentals among the stations has a very large support.

In order to reduce the computational requirements of LBQL, the lower and upper bounds updates are done every m steps and only if the difference between the current values of the bounds is greater than some threshold δ.

Since we can easily obtain inner DP results for all (s, a) each time the DP is solved, we perform the upper and lower bound updates for all (s, a) whenever an update is performed (as opposed to just at the current state-action pair). However, only the action-value of the current (s, a) is projected between the lower and upper bounds, so the algorithm is still asynchronous. The pseudo-code of LBQL with experience replay with these changes, is shown in Algorithm 3.

Algorithm 3 LBQL with Experience Replay (Full Details)

Input: Initial estimates $L_0 \leq Q_0 \leq U_0$, batch size K, stepsize rules $\alpha_n(s, a)$, $\beta_n(s, a)$, noise buffer $\mathcal{B}$ of size κ, number of steps between bound updates m, and threshold δ.
Output: Approximations $\{L_n\}$, $\{Q'_n\}$, and $\{U_n\}$.
Set $Q'_0 = Q_0$ and choose an initial state s_0.
for $n = 0, 1, 2, \ldots$ **do**
Choose an action a_n via some behavior policy (e.g., ϵ-greedy). Observe w_{n+1}.
Store w_{n+1} in $\mathcal{B}$.
Perform a standard Q-learning update:

$$Q_{n+1}(s_n, a_n) = Q'_n(s_n, a_n) + \alpha_n(s_n, a_n)\Big[r_n(s_n, a_n) + \gamma \max_a Q'_n(s_{n+1}, a) - Q'_n(s_n, a_n)\Big].$$

if $n \geq \kappa$ and $n \bmod m = 0$ and $|U_n(s_n, a_n) - L(s_n, a_n)| > \delta$ **then**
Sample randomly a batch $\mathcal{D} = \{w_1, w_2, \ldots, w_K\}$ and a sample path $\mathbf{w} = \{w_1, w_2, \ldots, w_\tau\}$ from $\mathcal{B}$, where $\tau \sim \text{Geom}(1 - \gamma)$.
Set $\varphi = Q_{n+1}$. Using $\mathbf{w}$ and $\mathcal{D}$, compute $\hat{Q}_0^U(s, a)$ and $\hat{Q}_0^L(s, a)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, using (9) and (10), respectively.
For all $(s, a) \in \mathcal{S} \times \mathcal{A}$, update upper and lower bounds:

$$U_{n+1}(s_n, a_n) = \Pi_{[-\rho, \infty]}\left[U_n(s_n, a_n) + \beta_n(s_n, a_n)\left[\hat{Q}_0^U(s_n, a_n) - U_n(s_n, a_n)\right]\right],$$
$$L_{n+1}(s_n, a_n) = \Pi_{[\infty, \rho]}\left[L_n(s_n, a_n) + \beta_n(s_n, a_n)\left[\hat{Q}_0^L(s_n, a_n) - L_n(s_n, a_n)\right]\right],$$

end if
Enforce upper and lower bounds:

$$Q'_{n+1}(s_n, a_n) = \Pi_{[L_{n+1}(s_n, a_n),\, U_{n+1}(s_n, a_n)]}\left[Q_{n+1}(s_n, a_n)\right]$$

end for

D Numerical Experiments Details

Let $\nu(s,a)$ and $\nu(s)$ be the number of times state-action pair (s,a) and state s, have been visited, respectively. For all algorithms, a polynomial learning rate $\alpha_n(s,a) = 1/\nu_n(s,a)^r$ is used, where $r = 0.5$. Polynomial learning rates have been shown to have a better performance than linear learning rates (van Hasselt, 2010).

We use a discount factor of $\gamma = 0.95$ for the pricing car-sharing/stormy gridworld problems, $\gamma = 0.9$ for the windy gridworld problem and $\gamma = 0.99$ for the repositioning problem. Moreover, we use an ϵ-greedy exploration strategy such that $\epsilon(s) = 1/\nu(s)^e$, where e is 0.4 for the four-stations pricing car-sharing problem and 0.5 for all the other problems. For the car-sharing/windy gridworld problems, the initial state-action values are chosen randomly such that $L_0(s,a) \leq Q_0(s,a) \leq U_0(s,a)$ where

$$L_0(s,a) = -R_{\max}/(1-\gamma) \quad \text{and} \quad U_0(s,a) = R_{\max}/(1-\gamma)$$

for all (s,a). For the stormy gridworld problem, we set the initial state-action values to zero (we find that a random initialization caused all algorithms except LBQL to perform extremely poorly).

We report LBQL parameters used in our numerical experiments in Table 1. Note that for a fair comparison, the parameter K of bias-corrected Q-learning algorithm is taken equal to K of LBQL in all experiments. In addition, the κ steps used to create the buffer for LBQL are included in the total number of steps taken. Results of the gridworld and car-sharing problems are averaged over 50 and 10 runs, respectively. All experiments were run on a 3.5 GHz Intel Xeon processor with 32 GB of RAM workstation.

Table 1: LBQL parameters.

	Parameter				
Problem	β	κ	K	m	δ
2-CS-R	0.01	40	20	10	0.01
2-CS	0.01	40	20	15	0.01
4-CS	0.01	1000	20	200	0.01
WG	0.2	100	10	10	0.01
SG	0.2	500	20	20	0.05

A detailed description of the environments is given in the next two sections.

D.1 Gridworld Examples

First we consider, windy gridworld, a well-known variant of the standard gridworld problem discussed in (Sutton & Barto, 2018). Then we introduce, stormy gridworld, a new environment that is more complicated than windy gridworld. The environments are summarized below.

Windy Gridworld. The environment is a 10×7 gridworld, with a cross wind pointing *upward*, (Sutton & Barto, 2018). The default *wind* values corresponding to each of the 10 columns are $\{0,0,0,1,1,1,2,2,1,0\}$. Allowable actions are {up, right, down, left}. If the agent happens to be in a column whose wind value is different from zero, the resulting next states are shifted upward by the "wind" whose intensity is stochastic, varying from the given values in each column by $\{-1,0,1\}$ with equal probability. Actions that corresponds to directions that takes the agent off the grid leave the location of the agent unchanged. The start and goal states are $(3,1)$ and $(3,8)$, respectively. The reward is -1 until the goal state is reached, where the reward is 0 thereafter.

Stormy Gridworld. Consider the stochastic windy gridworld environment. Now, however, we allow the wind to blow half the time as before and the other half it can blow from any of the three other directions. The horizontal wind values corresponding to each row from top to bottom are given by $\{0,0,1,1,1,1,0\}$. Also, it can randomly rain with equal probability in any of the central states that are more than two states away from the edges of the grid. The start and goal states are $(3,1)$ and $(3,10)$ respectively. Rain creates a puddle which affects the state itself and all of its neighboring states. The reward is as before except when the agent enters a puddle state the reward is -10.

D.2 Car-sharing Benchmark Examples

In this section, we give the detailed formulations of the two variants of the car-sharing benchmark, repositioning and pricing. The essential difference is that in the pricing version, the decision maker "repositions" by setting prices to induce directional demand (but does not have full control since this demand is random).

D.2.1 Repositioning Benchmark for Car-sharing

We consider the problem of repositioning cars for a two stations car-sharing platform, (He et al., 2019). The action is the number of cars to be repositioned from one station to the other, before random demand is realized. Since repositioning in both directions is never optimal, we use $r > 0$ to denote the repositioned vehicles from station 1 to 2 and $r < 0$ to denote repositioning from station 2 to 1. The stochastic demands at time t are $D_{1,t}$ and $D_{2,t}$ for stations 1 and 2 respectively, are i.i.d., discrete uniform, each supported on $\{3, \ldots, 9\}$. The rental prices are $p_1 = 3.5$ and $p_2 = 4$ for stations 1 and 2, respectively. All rentals are one-way (i.e., rentals from station 1 end up at station 2, and vice-versa). The goal is to maximize profit, where unmet demands are charged a lost sales cost $\rho_1 = \rho_2 = 2$ and repositioning cost $c_1 = 1$ for cars reposition from station 1 to 2 and $c_2 = 1.5$ for cars repositioned from station 2 to 1. We assume a total of $\bar{s} = 12$ cars in the system and formulate the problem as an MDP, with state $s_t \in \mathcal{S} = \{0, 1, \ldots, 12\}$ representing the number of cars at station 1 at beginning of period t. We denote by $V^*(s_t)$ the optimal value function starting from state s_t. The Bellman recursion is:

$$\begin{aligned} V^*(s_t) = \max_{s_t - \bar{s} \le r_t \le s_t} \mathbf{E}\Bigg[& \sum_{i \in \{1,2\}} p_i\, \omega_{it}(D_{i,t+1}) - \sum_{i \in \{1,2\}} \rho_i \Big(D_{i,t+1} - \omega_{it}(D_{i,t+1}) \Big) \\ & - c_1 \max(r_t, 0) + c_2 \min(r_t, 0) + \gamma V^*(s_{t+1}) \Bigg], \\ \omega_{1t}(D_{1,t+1}) &= \min(D_{1,t+1}, s_t - r_t), \\ \omega_{2t}(D_{2,t+1}) &= \min(D_{2,t+1}, \bar{s} - s_t + r_t), \\ s_{t+1} &= s_t - r_t + \omega_{2t}(D_{2,t+1}) - \omega_{1t}(D_{1,t+1}), \end{aligned} \tag{44}$$

where $\gamma \in (0, 1)$ is a discount factor. The repositioning problem for two stations is illustrated in Figure 6A. The nodes represent stations, solid arcs represent fulfilled demands, and dashed arcs represent repositioned vehicles.

D.2.2 Pricing Benchmark for Car-sharing

Suppose that a vehicle sharing manager is responsible for setting the rental price for the vehicles at the beginning of each period in an infinite planning horizon. We model a car sharing system with N stations. The goal is to optimize the prices to set for renting a car at each of the N stations; let the price at station i and time t be p_{it} for $i \in [N] := \{1, 2, \ldots, N\}$. Demands are nonnegative, independent and depends on the vehicle renting price according to a stochastic demand function

$$D_{it}(p_{it}, \epsilon_{i,t+1}) := \kappa_i(p_{it}) + \epsilon_{i,t+1},$$

where $D_{it}(p_{it}, \epsilon_{i,t+1})$ is the demand in period t, $\epsilon_{i,t+1}$ are random perturbations that are revealed at time $t+1$ and $\kappa_i(p_{it})$ is a deterministic demand function of the price p_{it} that is set at the beginning of period t at station $i \in [N]$. The random variables $\epsilon_{i,t+1}$ are independent with $\mathbf{E}[\epsilon_{i,t+1}] = 0$ without loss of generality. Furthermore, we assume that the expected demand $\mathbf{E}[D_{it}(p_{it}, \epsilon_{i,t+1})] = \kappa_i(p_{it}) < \infty$ is strictly decreasing in the rental price p_{it} which is restricted to a set of feasible price levels $[\underline{p}_i, \overline{p}_i]$ for all $i \in [N]$, where $\underline{p}_i, \overline{p}_i$ are the minimum and the maximum prices that can be set at station i, respectively. This assumption implies a one-to-one correspondence between the rental price p_{it} and the expected demand $d_{it} \in \mathfrak{D} := [\underline{d}_i, \overline{d}_i]$ for all $p_{it} \in [\underline{p}_i, \overline{p}_i]$ where $\underline{d}_i = \kappa_i(\overline{p}_i)$ and $\overline{d}_i = \kappa_i(\underline{p}_i)$.

The problem can be formulated as an MDP with state $\mathbf{s}_t$, which is a vector whose components represent the number of available cars at each of the N stations at beginning of period t. The state space is $\mathcal{S}^{N-1}$ with $\mathcal{S} = \{0, 1, \ldots, \bar{s}\}$ and $\bar{s}$ is the maximum number of cars in the vehicle sharing system. We assume that a customer at station i goes to station j with probability ϕ_{ij} for all $i, j \in [N]$. Let $Y_{ik,t+1}$ be a random variable taking values in $[N]$ that represents the random destination of customer k at station i, which is only observed at the beginning of period $t+1$. We have $Y_{ik,t+1} = j$ with probability ϕ_{ij}, so $Y_{ik,t+1}$ are i.i.d. for each customer k. Denote by l_{ij} the distance from station i to j, for all $i, j \in [N]$. We penalize unmet demands by a lost sales unit cost ρ_i, $i \in [N]$. The decision vector is $\mathbf{p}_t = \{p_{it} \in [\underline{p}_i, \overline{p}_i], \forall i \in [N]$.

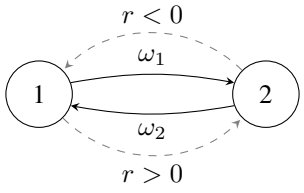
Let $V^*(\mathbf{s}_t)$ be the revenue-to-go function with number of available vehicles $\mathbf{s}_t$. Thus, we have the Bellman recursion

$$
\begin{aligned}
V^*(\mathbf{s}_t) &= \max_{\mathbf{p}_t} \mathbf{E}\Big[\sum_{i\in[N]} p_{it} \sum_{j\in[N]} l_{ij}\, \omega_{ijt}(\epsilon_{i,t+1}) - \sum_{i\in[N]} \rho_i\Big(\kappa_i(p_{it}) + \epsilon_{i,t+1} - \omega_{it}(\epsilon_{i,t+1})\Big) + \gamma V^*(\mathbf{s}_{t+1})\Big] \\
&\omega_{it}(\epsilon_{i,t+1}) = \min\left(\kappa_i(p_{it}) + \epsilon_{i,t+1}, s_{it}\right) \quad \forall\, i \in [N], \\
&\omega_{ijt}(\epsilon_{i,t+1}) = \sum_{k=1}^{\omega_{it}(\epsilon_{i,t+1})} \mathbb{1}_{\{Y_{ik,t+1}=j\}} \quad \forall\, i,j \in [N], \\
&s_{i,t+1} = s_{it} + \sum_{j\in[N]} \omega_{jit}(\epsilon_{i,t+1}) - \omega_{it}(\epsilon_{i,t+1}), \quad \forall\, i \in [N],
\end{aligned}
\tag{45}
$$

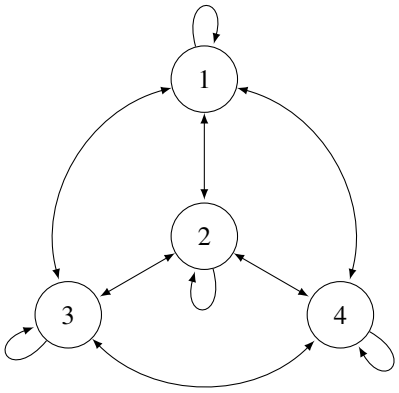
where $\gamma \in (0,1)$ is a discount factor. Note that the MDP in (45) can be reformulated using the action-value function $Q(\mathbf{s}_t, p_t)$ instead of $V(\mathbf{s}_t)$. The quantity $\omega_{it}(\epsilon_{i,t+1})$ represents the total fulfilled customer trips from station i at time t for a given realization of the noise $\epsilon_{i,t+1}$. Notice that in (45) there are two sources of randomness: the noise due to stochastic demand represented by ϵ_i, for all $i \in [N]$ and the noise due to the random distribution of fulfilled rentals between the stations, i.e., due to the random variables $Y_{i1}, \ldots, Y_{i\omega_{it}(\epsilon_{i,t+1})}$ for all $i \in [N]$. Due to the high dimensionality involved in the state, action, and noise spaces, solving (45) is computationally challenging.

Spatial Pricing in Two-Location Car-sharing. We first consider the pricing problem on two stations and 12 cars in total. The state space is $\mathcal{S} = \{0, 1, \ldots, 12\}$ representing the number of cars at station 1. All rentals are one-way. The prices, at each period t, are restricted to $p_{1t} \in [1, 6]$ and $p_{2t} \in [1, 7]$. The stochastic demand functions at period t are given by: $D_{1t}(p_{1t}, \epsilon_{1,t+1}) := 9 - p_{1t} + \epsilon_{1,t+1}$ and $D_{2t}(p_{2t}, \epsilon_{2,t+1}) := 10 - p_{2t} + \epsilon_{2,t+1}$ for stations 1 and 2 respectively. The random variables $\epsilon_{1,t+1}$ and $\epsilon_{2,t+1}$ are independent, discrete uniform, each supported on $\{-3, -2, \ldots, 3\}$. We use the discretized expected demands, as our actions: $d_{1t} \in \{3, \ldots, 8\}$ and $d_{2t} \in \{3, \ldots, 9\}$. The lost sales cost is 2 at both stations.

Spatial Pricing in Four-Location Car-sharing. Consider the car-sharing problem for four stations with $\bar{s} = 20$ cars and $d_{it} \in \{3, 4\}$ for each station. In total there are 1771 states and 16 actions. The random variables $\epsilon_{i,t+1}$ are independent, discrete uniform, each supported on $\{-3, -2, \ldots, 3\}$. We consider both one way and return trips at each station. Figure 6B shows an illustration of the stations (nodes) and the rentals between the stations (arcs). The probabilities $\phi_{ij} = 0.25$ for all $i, j \in \{1, 2, 3, 4\}$ and the lost sales costs (ρ_i) are $1.7, 1.2, 1.5, 2$ at stations $1, 2, 3, 4$, respectively. The distance between the stations are taken such that $l_{ij} = 1$ if $i = j$, and the other distances being symmetrical, meaning $l_{ij} = l_{ji}$ with $l_{12} = 1.8$, $l_{13} = 1.5$, $l_{14} = 1.4$, $l_{23} = 1.6$, $l_{24} = 1.1$, and $l_{34} = 1.2$.



(A) Repositioning problem with 2 stations.



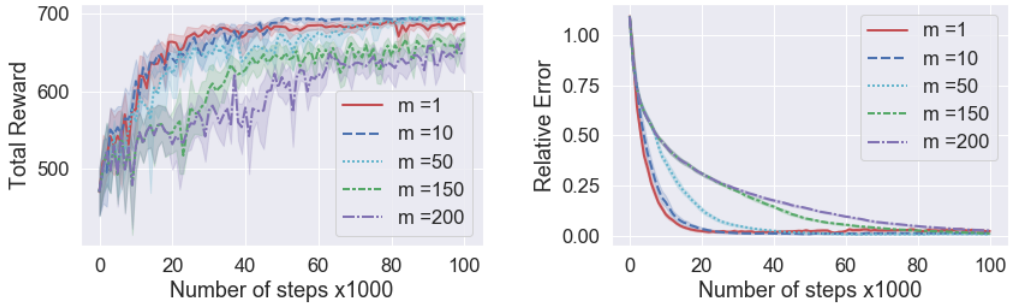
(B) Pricing problem with 4 stations.

Figure 6: Illustrations of the repositioning and pricing car-sharing problems.

D.3 Sensitivity Analysis

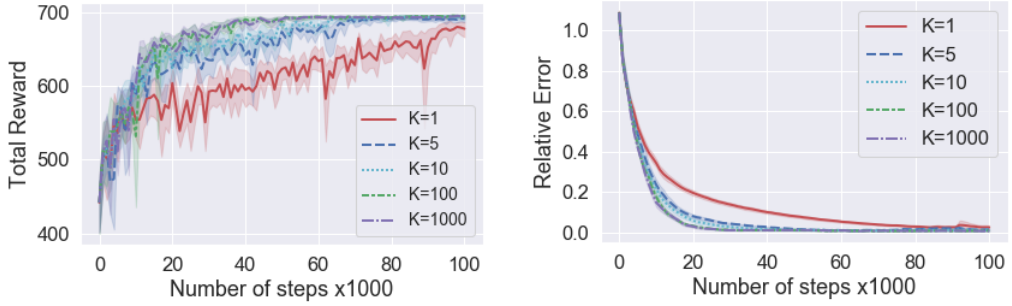
We also perform sensitivity analysis on the five algorithms with respect to the learning rate and exploration parameters r and e for the car-sharing problem with two stations. Here, r controls the polynomial learning rate defined by, $\alpha_n(s,a) = 1/\nu_n(s,a)^r$ and e controls the ϵ-greedy exploration strategy, where ϵ is annealed according to $\epsilon(s) = 1/\nu(s)^e$. We use $\nu(s,a)$ and $\nu(s)$ to denote the number of times a state-action pair (s,a) and state s, have been visited, respectively. We report our results in Table 2. These results show the average number of iterations and CPU time until each algorithm first reach 50%, 20%, 5%, 1% relative error for each case of the parameters e and r while keeping all other parameters as before. The "-" indicates that the corresponding % relative error for the corresponding case was not achieved during the course of training. The values in the table are obtained by averaging five independent runs for each case. Except for the few cases where BCQL performs slightly better, LBQL once again drastically outperforms the other algorithms and exhibits *robustness* against the learning rate and exploration parameters, an important practical property that the other algorithms seem to lack.

The effect of varying parameters m and K of LBQL is presented in Figure 7. These plots are obtained by tuning parameters $m \in \{1, 10, 50, 150, 200\}$ and $K \in \{1, 5, 10, 100, 1000\}$ of LBQL algorithm in the car-sharing problem with two stations. All other parameters are kept the same as before. Figures 7A and 7C show the mean total reward with a 95% CI. Figures 7B and 7D show the mean and 95% CI of the relative error given by: $\|V_n - V^*\|_2/\|V^*\|_2$. The results are obtained from 10 independent runs. Using larger values of m reduces the strength of LBQL in both the performance and relative error metrics, as shown in Figures 7A and 7B. This is expected since the effect of the bounds fades as we update the bounds less frequently. Interestingly, we can see from the performance plot that $m = 10$ strikes a good balance between how often to do the bounds and Q-learning updates and achieves a performance that is slightly better and more stable than that of $m = 1$ after about half of the training process (50,000 steps). In terms of the sample size K, Figures 7C and 7D clearly show that larger values of K improve the performance of LBQL in terms of performance and relative error measures. This is not unexpected because a larger sample yields a better approximation of the penalty.



(A) Performance (2-CS)

(B) Relative Error (2-CS)



(C) Performance (2-CS)

(D) Relative Error (2-CS)

Figure 7: Plots showing the effect of tuning the parameters m and K of LBQL algorithm.

Table 2: Computational results for different exploration & learning rate parameters. Bold numbers indicate the best performing algorithm.

	e	r	% Relative error 50% n	50% t (s)	20% n	20% t (s)	5% n	5% t (s)	1% n	1% t (s)
LBQL	0.4	0.5	3,672.6	1.9	**9,323.6**	**4.6**	**18,456.6**	**8.9**	**33,054.0**	**15.7**
		0.6	**3,632.0**	1.7	**9,147.4**	**4.4**	**18,270.0**	**8.6**	**39,624.8**	**18.6**
		0.7	**3,725.2**	**1.8**	**9,087.4**	**4.3**	**18,217.8**	**8.6**	**41,941.2**	**19.7**
		0.8	**3,698.2**	**1.8**	**9,321.8**	**4.4**	**20,860.0**	**9.9**	**53,752.6**	**25.6**
		0.9	**3,992.2**	**1.9**	**10,119.2**	**4.9**	**23,070.2**	**11.0**	**80,252.8**	**38.4**
	0.5	0.5	3,316.0	1.6	**8,040.2**	**3.8**	**15,050.2**	**7.2**	**27,912.8**	**13.2**
		0.6	3,514.4	1.7	**8,529.4**	**4.0**	**16,595.6**	**7.9**	**36,100.2**	**17.0**
		0.7	**3,531.8**	**1.7**	**8,712.8**	**4.1**	**17,835.6**	**8.6**	**46,010.0**	**21.9**
		0.8	**3,449.2**	**1.6**	**8,571.8**	**4.0**	**18,152.6**	**8.5**	**65,007.6**	**30.4**
		0.9	**3,398.4**	**1.6**	**8,346.4**	**3.9**	**18,844.8**	**8.8**	**99,820.2**	**46.6**
	0.6	0.5	2,877.4	1.3	**7,129.0**	**3.3**	**13,046.0**	**6.2**	**23,822.0**	**11.2**
		0.6	3,182.8	1.5	**8,066.0**	**3.8**	**15,421.4**	**7.3**	**33,286.0**	**15.6**
		0.7	**2,979.4**	1.4	**7,625.6**	**3.6**	**15,414.6**	**7.2**	**34,238.0**	**15.9**
		0.8	**3,272.6**	**1.6**	**8,431.0**	**4.1**	**17,809.2**	**8.5**	**114,032.8**	**54.2**
		0.9	**3,185.6**	**1.5**	**8,480.0**	**4.1**	**19,242.4**	**9.2**	**123,331.2**	**58.8**
BCQL	0.4	0.5	**3,200.8**	**1.0**	22,455.0	7.0	65,329.0	20.3	107,785.6	33.7
		0.6	4,618.2	**1.5**	43,724.6	13.6	159,662.6	49.6	292,421.0	34.9
		0.7	8,059.4	2.5	123,484.2	38.4	-	-	-	-
		0.8	17,287.0	5.3	-	-	-	-	-	-
		0.9	67,162.2	20.9	-	-	-	-	-	-
	0.5	0.5	**2,209.6**	**0.7**	15,604.0	4.9	48,715.6	15.3	80,317.4	25.2
		0.6	**3,274.2**	**1.0**	31,422.8	9.8	124,319.6	38.7	243,101.4	75.6
		0.7	5,619.6	1.8	89,857.0	27.8	-	-	-	-
		0.8	11,417.0	3.6	-	-	-	-	-	-
		0.9	42,605.4	13.1	-	-	-	-	-	-
	0.6	0.5	**1,830.4**	**0.6**	11,639.6	3.6	35,763.0	11.1	61,249.0	19.0
		0.6	**2,612.4**	**0.8**	23,571.6	7.4	92,101.4	28.8	177,127.6	55.5
		0.7	4,371.2	**1.3**	66,526.0	20.5	-	-	-	-
		0.8	9,028.2	2.8	297,368.6	17.9	-	-	-	-
		0.9	31,673.6	9.8	-	-	-	-	-	-
SQL	0.4	0.5	7,750.6	1.9	37,889.8	9.1	93,820.0	22.5	141,171.0	33.8
		0.6	11,329.4	2.8	75,364.0	18.3	233,422.0	56.4	-	-
		0.7	20,131.4	4.8	212,767.0	51.0	-	-	-	-
		0.8	46,986.8	11.3	-	-	-	-	-	-
		0.9	182,890.0	43.8	-	-	-	-	-	-
	0.5	0.5	6,122.2	1.5	30,944.8	7.4	79,167.4	19.0	120,527.8	29.0
		0.6	9,166.6	2.2	62,540.6	14.9	201,822.4	48.2	-	-
		0.7	15,835.6	3.8	174,233.6	42.0	-	-	-	-
		0.8	36,548.8	8.7	-	-	-	-	-	-
		0.9	157,029.0	37.7	-	-	-	-	-	-
	0.6	0.5	4,984.0	1.2	24,989.0	6.0	64,605.4	15.4	98,554.6	23.6
		0.6	7,396.2	1.8	50,282.4	12.0	165,574.2	39.8	-	-
		0.7	13,018.8	3.1	143,142.6	34.1	-	-	-	-
		0.8	29,201.0	6.9	-	-	-	-	-	-
		0.9	122,335.6	29.2	-	-	-	-	-	-

(continued on next page)

Table 2: *(continued)*

	e	r	% Relative error 50% n	50% t (s)	20% n	20% t (s)	5% n	5% t (s)	1% n	1% t (s)
QL	0.4	0.5	7,743.0	1.7	38,114.2	8.2	93,303.4	20.2	136,851.4	29.6
		0.6	11,644.0	2.5	76,625.0	16.6	232,679.4	50.9	-	-
		0.7	20,181.6	4.4	212,401.4	46.3	-	-	-	-
		0.8	45,987.2	10.1	-	-	-	-	-	-
		0.9	191,442.2	41.9	-	-	-	-	-	-
	0.5	0.5	6,143.6	1.3	30,996.2	6.8	78,131.8	16.9	116,361.2	25.3
		0.6	9,331.6	2.0	63,998.2	13.9	204,593.6	44.6	-	-
		0.7	16,247.0	3.5	178,842.8	38.4	-	-	-	-
		0.8	38,297.0	8.2	-	-	-	-	-	-
		0.9	165,835.8	35.7	-	-	-	-	-	-
	0.6	0.5	5,005.2	1.1	24,877.2	5.4	63,777.8	13.7	96,402.0	20.8
		0.6	7,547.0	1.9	51,369.4	13.1	166,179.2	42.3	289,882.8	46.1
		0.7	13,288.2	3.1	144,318.2	33.1	-	-	-	-
		0.8	30,172.6	6.5	-	-	-	-	-	-
		0.9	139,952.6	30.3	-	-	-	-	-	-
Double-QL	0.4	0.5	224,490.2	51.2	-	-	-	-	-	-
		0.6	-	-	-	-	-	-	-	-
		0.7	-	-	-	-	-	-	-	-
		0.8	-	-	-	-	-	-	-	-
		0.9	-	-	-	-	-	-	-	-
	0.5	0.5	-	-	-	-	-	-	-	-
		0.6	-	-	-	-	-	-	-	-
		0.7	-	-	-	-	-	-	-	-
		0.8	-	-	-	-	-	-	-	-
		0.9	-	-	-	-	-	-	-	-
	0.6	0.5	-	-	-	-	-	-	-	-
		0.6	-	-	-	-	-	-	-	-
		0.7	-	-	-	-	-	-	-	-
		0.8	-	-	-	-	-	-	-	-
		0.9	-	-	-	-	-	-	-	-