

## Appendices

### A PROOF FOR LOWER BOUNDS (THEOREM 2)

We construct here an easy instance of the episodic inventory control problem (as in Section 6), for which the regret of any algorithm must be at least  $\Omega(\sqrt{HT})$ .

*Proof.* Suppose for any time  $h$  in an episode, the demand distribution is  $h+100$  units w.p.  $0.5 + \frac{1}{\sqrt{K}}$ , and  $h+200$  units w.p.  $0.5 - \frac{1}{\sqrt{K}}$ . Note that this is not a constant gap, because  $K = \Theta(T)$ . Suppose the unit costs for holding, backlogging, and lost-sales penalty are all the same. We generously allow the algorithm to have the correct prior that the best base stock level is one of these two actions, and the other actions are worse than these two actions. Then for each time step  $h$ , our problem of estimating the  $Q$ -values degenerates to the stochastic full-feedback online bandit problem.

It is a well-known result that in this case, each stage  $h$  will incur at least a  $\Omega(\sqrt{K})$  regret across the  $K$  episodes. In particular, at any time step of any episode, the probability of any algorithm choosing the wrong action is lower-bounded by  $\frac{1}{12}$ : see Corollary 2.10 in Slivkins (2019). Then at each time step, the algorithm incur a  $\Omega(\frac{1}{12\sqrt{K}})$  expected regret. This regret at stage  $h$  across the  $K$  episodes sum up to  $\Omega(\sqrt{K})$  expected regret. Since there are  $H$  time steps with demand independent from each other, we have that the regret of this example is lower bounded by  $\Omega(H\sqrt{K}) = \Omega(\sqrt{HT})$  regret. Note that even though we assume the algorithm receives full information feedback at each time step, Corollary 2.10 in Slivkins (2019) still applies by scaling the time horizon by a factor of 2, which does not affect the regret bound. Then we put back the  $\Theta(M \cdot \max(|o_h|, |b_h|))$  factor (or  $\Theta(M \cdot \max(|o_h|, |p_h|))$  factor) because in the Preliminaries we scaled the unit costs down by  $\Theta(M)$  to have the reward for each time period bounded by 1.  $\square$

### B MISSING PROOFS FOR $HQL$

*Proof. (Lemma 3)* We prove number (4) by induction. For the base case  $t = 1$ , we have  $\sum_{i=1}^t \frac{\alpha_t^i}{\sqrt{i}} = \alpha_1^1 = 1$  so the statement holds. For  $t \geq 2$ , by the relationship  $\alpha_t^i = (1 - \alpha_t)\alpha_{t-1}^i$  for  $i = 1, \dots, t-1$  we have

$$\sum_{i=1}^t \frac{\alpha_t^i}{\sqrt{i}} = \frac{\alpha_t}{\sqrt{t}} + (1 - \alpha_t) \sum_{i=1}^{t-1} \frac{\alpha_{t-1}^i}{\sqrt{i}} \quad (7)$$

Assuming the inductive hypothesis holds, on the one hand,

$$\frac{\alpha_t}{\sqrt{t}} + (1 - \alpha_t) \sum_{i=1}^{t-1} \frac{\alpha_{t-1}^i}{\sqrt{i}} \geq \frac{\alpha_t}{\sqrt{t}} + \frac{1 - \alpha_t}{\sqrt{t-1}} \geq \frac{\alpha_t}{\sqrt{t}} + \frac{1 - \alpha_t}{\sqrt{t}} = \frac{1}{\sqrt{t}} \quad (8)$$

where the first inequality holds by the inductive hypothesis. On the other hand,

$$\begin{aligned} \frac{\alpha_t}{\sqrt{t}} + (1 - \alpha_t) \sum_{i=1}^{t-1} \frac{\alpha_{t-1}^i}{\sqrt{i}} &\leq \frac{\alpha_t}{\sqrt{t}} + \frac{(1 + 1/H)(1 - \alpha_t)}{\sqrt{t-1}} = \frac{H+1}{\sqrt{t}(H+t)} + \frac{(1 + 1/H)\sqrt{t-1}}{H+t} \\ &\leq \frac{H+1}{\sqrt{t}(H+t)} + \frac{(1 + 1/H)\sqrt{t}}{H+t} \leq \frac{(1 + 1/H)}{\sqrt{t}} \end{aligned} \quad (9)$$

where the first inequality holds by the inductive hypothesis.  $\square$

This is a tighter bound than the bound in Jin et al. (2018). For rest of the lemma, see Lemma 4.1 in Jin et al. (2018).

The following proof for shortfall decomposition is adapted from Benjamin Van Roy's reinforcement learning notes for the class MS 338 at Stanford University.

*Proof. (Lemma 4)* For any policy  $\pi$ , let  $y_h^k$  denote the action the policy  $\pi_k$  takes at stage  $h$  of episode  $k$ . Let  $R_h$  denote the expected reward of  $y_h^k$ .

$$\mathbb{E}_\pi [Q^*(x_h^k, y_h^k)] = \mathbb{E}_\pi [Z_{h+1}] \quad \text{where } Z_{h+1} = \begin{cases} R_h + \max_y Q^*(x_{h+1}^k, y) & \text{if } h < H \\ R_h & \text{if } h = H \end{cases}$$

Therefore,

$$\begin{aligned} V_1^* - V_1^{\pi_k} &= \mathbb{E}_\pi \left[ \max_{a \in \mathcal{A}} Q^*(x_1^k, a) - \sum_{h=1}^H R_h \right] \\ &= \mathbb{E}_\pi \left[ \max_{a \in \mathcal{A}} Q^*(x_1^k, a) - \sum_{h=1}^H (R_h - Z_{h+1} + Q^*(x_h^k, y_h^k)) \right] \\ &= \mathbb{E}_\pi \left[ \sum_{h=1}^H (\max_{a \in \mathcal{A}} Q^*(x_h^k, a) - Q^*(x_h^k, y_h^k)) \right] \end{aligned}$$

□

*Proof. (Lemma 5)* From the Bellman optimality equation (3), and the fact that  $\sum_{i=0}^{k-1} \alpha_{k-1}^i = 1$ , we have

$$Q_h^*(y) = \alpha_{k-1}^0 Q_h^*(y) + \sum_{i=1}^{k-1} \alpha_{k-1}^i \left[ \mathbb{E}_{x', \tau_h^i(y) \sim \mathbb{P}(\cdot | x, y)} [\tilde{r}_{\tau_h^i(x, y)}^*(y) + V_{\tau_h^i(x, y)}^*(x'_{\tau_h^i(y)})] \right]$$

Subtracting Equation 5 from this equation, and adding some of the middle terms that cancel with themselves gives us Lemma 5. □

*Proof. (Lemma 6)* Since we assume that given a fixed value  $D_h$ , the next state  $x_{h+1}(y_h)$  is increasing in  $y_h$ , and  $a_h(x_h)$  is increasing in  $x_h$  for the lower one-sided-feedback problem, we conclude that the (deterministic given  $D_h$ ) dynamics are monotone with respect to any simulation starting point  $x_h$ . Since the algorithm chooses at least the maximal action in  $A_h^k$  at all times, this implies it can observe the simulated trajectory started from any  $x_h \in A_h^k$  for any  $k, h \in [K] \times [H]$ .

Let  $\mathcal{F}_h^i$  be the  $\sigma$ -field generated by all the random variables until episode  $i$ , stage  $h$ . Then for any  $\tau \in [K]$ ,  $\left( V_{\tau_h^i(x, y)}^*(x_{\tau_h^i(x, y)}^i) + \tilde{r}_{\tau_h^i(x, y)}^* - \mathbb{E}_{\tilde{r}^*, x', \tau_h^i(x, y) \sim \mathbb{P}(\cdot | x, y)} [\tilde{r}_{\tau_h^i(x, y)}^* + V_{\tau_h^i(x, y)}^*(x'_{\tau_h^i(x, y)})] \right)_{i=1}^\tau$  is a martingale difference sequence w.r.t. the filtration  $\{\mathcal{F}_h^i\}_{i \geq 0}$ . Then by Azuma-Hoeffding Theorem, we have that with probability at least  $1 - (1/AT)^9$ :

$$\begin{aligned} & \left| \sum_{i=1}^{k-1} \alpha_{k-1}^i \cdot \left( V_{\tau_h^i(x, y)}^*(x_{\tau_h^i(x, y)}^i) + \tilde{r}_{\tau_h^i(x, y)}^* - \mathbb{E}_{\tilde{r}^*, x', \tau_h^i(x, y) \sim \mathbb{P}(\cdot | x, y)} [\tilde{r}_{\tau_h^i(x, y)}^* + V_{\tau_h^i(x, y)}^*(x'_{\tau_h^i(x, y)})] \right) \right| \\ & \leq \frac{cH}{2} \sqrt{\sum_{i=1}^{k-1} (\alpha_{k-1}^i)^2 \cdot \iota} \leq c \sqrt{\frac{H^3 \iota}{k-1}} \quad \text{for any constant } c \geq 2\sqrt{2}. \end{aligned} \tag{10}$$

By union bound, we have with probability at least  $1 - (1/AT)^8$  that for any  $x, h, k, y \in A_h^k$ ,

$$\begin{aligned} & \left| \sum_{i=1}^{k-1} \alpha_{k-1}^i \left( V_{\tau_h^i(x, y)}^*(x_{\tau_h^i(x, y)}^i) + \tilde{r}_{\tau_h^i(x, y)}^* - \mathbb{E}_{\tilde{r}^*, x', \tau_h^i(x, y) \sim \mathbb{P}(\cdot | x, y)} [\tilde{r}_{\tau_h^i(x, y)}^* + V_{\tau_h^i(x, y)}^*(x'_{\tau_h^i(x, y)})] \right) \right| \\ & \leq c \sqrt{\frac{H^3 \iota}{k-1}} \end{aligned}$$

Then Lemma 6 follows immediately this equation and Lemma 5. □

*Proof. (Upper Bound on  $\delta_h$ 's)* We set  $d_h = (\delta_h) \cdot (1 + \frac{1}{H})^h$  and observe that the recurrence implies

$$d_h = d_{h+1} + H + 2\sqrt{2}\sqrt{H^3\iota} \quad (11)$$

Then from this recursion we see  $d_h \leq H^2 + 2\sqrt{2H^5\iota}$  for all  $h$ . Since  $d_h, \delta_h$  differ by a constant factor  $(1 + \frac{1}{H})^h$ , we have  $\delta_h = \frac{H^2 + 2\sqrt{2H^5\iota}}{(1 + \frac{1}{H})^h} \leq 4\sqrt{H^5\iota}$ .  $\square$

*Proof. (Lemma 7)* We prove by backward induction. Note that all of our statements below hold with high probability. In particular, we will use Azuma-Hoeffding no more than  $AT$  times in the below, with each use holding with probability at least  $1/(AT)^9$ . Under the assumption that each use of Azuma-Hoeffding holds we will obtain the statement of the Lemma. Our proof goes by induction; for the base case  $\delta_{H+1} = 0$  satisfies the Inequality 7 (actually equality here) with probability 1 based on Bellman equations.

Now suppose inequality 7 is true for any  $k \in [K]$ ,  $x \in \mathcal{S}$ , for any  $h' = \tau_h^k(x, a)$  that has  $a \in A_h^k$ :

$$\max_{y \in A_{\tau_h^k(x, a)}^k} |(Q_{\tau_h^k(x, a)}^k - Q_{\tau_h^k(x, a)}^*)(y)| \leq \frac{\delta_{\tau_h^k(x, a)}}{\sqrt{k-1}}, \forall a \in A_h^k, \text{ w.h.p.}$$

Now we induct on the previous stage  $h' = h$ . By Lemma 6, with probability at least  $1 - 1/(AT)^8$

$$\begin{aligned} \max_{y \in A_h^k} |(Q_h^k - Q_h^*)(y)| &\leq \max_{a \in A_h^k} \left\{ \alpha_{k-1}^0 H \right. \\ &\quad \left. + \sum_{i=1}^{k-1} \alpha_{k-1}^i \left[ \left( V_{\tau_h^i(x, a)}^{i+1} - V_{\tau_h^i(x, a)}^* \right) \left( x_{\tau_h^i(x, a)}^i \right)' + \tilde{r}_{h, \tau_h^i(x, a)}^i - \tilde{r}_{h, \tau_h^i(x, a)}^* \right] + c\sqrt{\frac{H^3\iota}{k-1}} \right\} \end{aligned}$$

Since based on our inductive hypothesis, we have

$$\begin{aligned} \max_{a \in A_h^k} \left[ \left( V_{\tau_h^i(x, a)}^{i+1} - V_{\tau_h^i(x, a)}^* \right) \left( x_{\tau_h^i(x, a)}^i \right)' + \tilde{r}_{h, \tau_h^i(x, a)}^i - \tilde{r}_{h, \tau_h^i(x, a)}^* \right] \\ \leq \max_{y \in A_{\tau_h^i(x, a)}^i} |(Q_{\tau_h^i(x, a)}^{i+1} - Q_{\tau_h^i(x, a)}^*)(y)| \leq \frac{\delta_{\tau_h^i(x, a)}}{\sqrt{i}} \end{aligned}$$

then

$$\max_{y \in A_h^k} |(Q_h^k - Q_h^*)(y)| \leq \max_{a \in A_h^k} \left\{ \alpha_{k-1}^0 H + \left( \sum_{i=1}^{k-1} \alpha_{k-1}^i \cdot \frac{\delta_{\tau_h^i(x, a)}}{\sqrt{i}} \right) + c\sqrt{\frac{H^3\iota}{k-1}} \right\}. \quad (12)$$

We can bound  $\alpha_{k-1}^0$  by  $\frac{1}{\sqrt{k}}$ , and bound  $\sum_{i=1}^{k-1} \alpha_{k-1}^i \cdot \frac{\delta_{\tau_h^i(x, a)}}{\sqrt{i}}$  by  $\frac{1+1/H}{\sqrt{k-1}} \delta_{\tau_h^i(x, a)}$  using Lemma 3:

$$\begin{aligned} \max_{y \in A_h^k} |(Q_h^k - Q_h^*)(y)| &\leq \frac{1}{\sqrt{k}} H + \frac{1+1/H}{\sqrt{k-1}} \delta_{\tau_h^i(x, a)} + c\sqrt{\frac{H^3\iota}{k}} \\ &\leq \frac{1}{\sqrt{k-1}} H + \frac{1+1/H}{\sqrt{k-1}} \delta_{h+1} + c\sqrt{\frac{H^3\iota}{k-1}} = \frac{\delta_h}{\sqrt{k-1}} \end{aligned} \quad (13)$$

where the second inequality is because  $\tau_h^i(x, a) \geq h+1$  and  $\delta_h$ 's is a decreasing sequence. The last equality is true based on the recursive definition of  $\delta_h$ .  $\square$

*Proof. (Lemma 8)* Recall for any  $(x, h, k)$ ,  $y_h^{k*} = \arg \max_{y \in A_h^k} Q_h^k(y)$  in  $HQL$ . Suppose  $y_h^* \notin A_h^k$ , then  $Q_h^k(y_h^*) < Q_h^k(y_h^{k*}) - \frac{8\sqrt{H^5\iota}}{\sqrt{k-1}} = Q_h^k(x, y_h^{k*}) - \frac{2\delta_h}{\sqrt{k-1}}$ . Then we need either  $Q_h^k(y_h^*) < Q_h^*(y_h^*) - \frac{\delta_h}{\sqrt{k-1}}$  or  $Q_h^k(y_h^{k*}) > Q_h^*(y_h^{k*}) + \frac{\delta_h}{\sqrt{k-1}}$ . Thus by Lemma 7,  $\text{Prob}(y_h^* \notin A_h^k(x)) \leq \frac{1}{(AT)^5}$ .  $\square$

*Proof. (Lemma 9)* Lemma 7 says for any  $y \in A_h^k$ , our estimated  $Q_h^k(y)$  differs from the optimal value  $Q_h^*(y)$  by at most  $\frac{\delta_h}{\sqrt{k-1}}$  with high probability at least  $1 - \frac{1}{(AT)^5}$ . Therefore, the optimal Q-value of the optimal policy's action  $Q^*(y_h^*)$  is at most  $\frac{\delta_h}{\sqrt{k-1}}$  more than the estimated Q-value of our estimated best arm  $Q_h^k(y_h^{k*})$ , with high probability at least  $1 - \frac{1}{(AT)^5}$ . Any action we take in  $A_h^k$  has an estimated Q-value no more than  $\frac{8\sqrt{H^5\epsilon}}{\sqrt{k-1}} = \frac{2\delta_h}{\sqrt{k-1}}$  lower than  $Q_h^k(y_h^{k*})$  base on our algorithm. Therefore, the optimal Q-value of the optimal policy's action  $Q^*(y_h^*)$  is at most  $\frac{3\delta_h}{\sqrt{k-1}}$  more than the estimated Q-value of any action  $y \in A_h^k(x)$ , with high probability at least  $1 - \frac{1}{(AT)^5}$ . Then again, by Lemma 7, we know that the optimal Q-value of the optimal policy's action  $Q^*(y_h^*)$  is at most  $\frac{4\delta_h}{\sqrt{k-1}}$  more than the optimal Q-value of any action in  $A_h^k$ , with high probability at least  $1 - \frac{2}{(AT)^5}$ .  $\square$

*Proof. (Lemma 10)* From Lemma 8, we know that with probability at least  $1 - \frac{1}{(AT)^5}$ , the optimal action is in the running set, which is inaccessible. Then recall the assumptions that the value functions are concave and that the feasible action set at any time is an interval of the form  $\mathcal{A} \cap [a, \infty)$  for some  $a$  dependent on the state. So if we cannot play in the running set, then the running set, and hence w.h.p. the true optimal action, is contained in  $(-\infty, a)$ . By concavity, this implies that the closest feasible action to the running set is optimal in this case w.p. at least  $1 - \frac{1}{(AT)^5}$ .  $\square$

## C MISSING PROOFS FOR INVENTORY CONTROL

We gave a more detailed description of the backlogged model and the lost-sales model of the episodic stochastic inventory control problems.

**Lemma 11.** *For any  $h \in [H]$ , the optimal  $V$ -value function  $V_h^*(x)$  is concave in  $x$ , and the optimal  $Q$ -value function  $Q_h^*(y)$  is concave in  $y$ . This is true for the lost sales and the backlogged models.*

*Proof. (Lemma 11)* We prove this by backward induction. The base case is  $Q_H^*(x, y)$  and  $V_H^*(x)$ . Since  $Q_H^*(y)$  is just the expectation of a one time reward for the last period, we know that it is  $Q_H^*(x, y) = r_H(x, y, D_H) = -[o_H(y - D_H)^+ + p_H \min(y, D_H)]$ . This function is obviously concave in  $y$ . Note that the Q-values are not affected by  $x$  for the inventory control problems. Since  $V_H^*(x) = \max_{y \geq x} Q_H^*(x, y)$ , the graph of  $V_H^*(x)$  is constant on the left side of  $x = \arg \max_y Q_H^*(x, y)$ , and then goes down with a slope of  $-o_H$  on the right side of  $x = \arg \max_y Q_H^*(x, y)$ . So  $V_H^*(x)$  is obviously also concave.

Now suppose  $Q_{h+1}^*(x, y)$  and  $V_{h+1}^*(x)$  are concave. It remains to show concavity of  $Q_h^*(x, y)$  and  $V_h^*(x)$ .

We know  $Q_h^*(x, y) = \mathbb{E}[V_{h+1}^*(y - D_h) + r_h(x, y, D_h)]$ . We know  $r_h(x, y, D_h)$  is concave in  $y$  for the same reason that  $Q_H^*(x, y)$  is concave. We know that  $V_{h+1}^*(x)$  is concave in  $x$  from our induction hypothesis, which means  $V_{h+1}^*(y - D_h)$  is concave in  $y$  for any value of  $D_h$ . Therefore,  $\mathbb{E}[V_{h+1}^*(y - D_h) + r_h]$  is also concave, being a weighted average of concave functions. So we know  $Q_h^*(x, y)$  is also concave in  $y$ . Then again  $V_h^*(x) = \max_{y \geq x} Q_h^*(x, y)$  is concave for the same reason why  $V_H^*(x)$  is concave.  $\square$

*Proof. (Assumption of 0 Purchasing Costs)* We want to show that for the episodic lost-sales (and similarly for the backlogged) model, we can amortize the unit purchasing costs  $c_h$  into unit holding costs  $o_h$  and unit lost-sales penalty  $p_h$ , so that without loss of generality we can assume 0 purchasing costs.

$$\forall h \geq 2, y_h - x_h = y_h - D_h + D_h - x_h = (y_h - D_h)^+ - (D_h - y_h)^+ + D_h - (y_{t-1} - D_{t-1})^+ \quad (14)$$

Let  $c_h$  denote the unit purchasing cost, then the total sum of costs starting from stage 2 is

$$\begin{aligned} & \sum_{h=2}^H \left( c_h(y_h - x_h) + o_h(y_h - D_h)^+ + p_h(D_h - y_h)^+ \right) \\ &= \sum_{h=2}^H \left( c_h D_h - c_h(y_{h-1} - D_{h-1})^+ + (o_h + c_h)(y_h - D_h)^+ + (p_h - c_h)(D_h - y_h)^+ \right) \end{aligned}$$

And the cost of stage 1 is equal to  $o_1(y_1 - D_1)^+ + p_1(D_1 - y_1)^+ + c_1((y_1 - D_1)^+ - (D_1 - y_1)^+ + D_1 - x_1)$ .

Let  $c_{H+1} \geq 0$  denote the salvage price at which we sell the remaining inventory  $(y_H - D_H)^+$  at the end of each episode. Then the total sum of costs from stage 1 to H is

$$\begin{aligned} & \sum_{h=2}^H \left( c_h D_h - c_h(y_{h-1} - D_{h-1})^+ + (o_h + c_h)(y_h - D_h)^+ + (p_h - c_h)(D_h - y_h)^+ \right) \\ &+ c_1(y_1 - D_1)^+ - c_1(D_1 - y_1)^+ + c_1 D_1 - c_1 x_1 + o_1(y_1 - D_1)^+ + p_1(D_1 - y_1)^+ - c_{H+1}(y_H - D_H)^+ \\ &= \sum_{h=2}^H \left( c_h D_h - c_h(y_{h-1} - D_{h-1})^+ + (o_h + c_h)(y_h - D_h)^+ + (p_h - c_h)(D_h - y_h)^+ \right) \\ &+ c_1(y_1 - D_1)^+ - c_1(D_1 - y_1)^+ + c_1 D_1 - c_1 x_1 + o_1(y_1 - D_1)^+ + p_1(D_1 - y_1)^+ - c_{H+1}(y_H - D_H)^+ \\ &= \sum_{h=1}^H c_h D_h + \sum_{h=1}^H \left( (o_h + c_h - c_{h+1})(y_h - D_h)^+ + (p_h - c_h)(D_h - y_h)^+ \right) - c_1 x_1 \end{aligned}$$

Since  $\sum_{h=1}^H c_h D_h$  and  $-c_1 x_1$  are fixed costs independent of our action, we can take them out of our consideration. Then we can effectively consider the cost of each stage  $h$  is just  $o'_h(y_h - D_h)^+ + p'_h(D_h - y_h)^+$ , where  $o'_h = o_h + c_h - c_{h+1}$  is the adjusted holding cost, and  $p'_h = p_h - c_h$  is the adjusted lost-sales penalty.  $\square$

## D COMPARISON WITH EXISTING Q-LEARNING ALGORITHMS

For Jin et al. (2018), suppose we discretize the state and action space optimally with step-size  $\epsilon_1$  to apply Jin et al. (2018) to the backlogged/lost-sales episodic inventory control problem with continuous action and state space. Then the  $\text{Regret}_{gap}$  we get is  $\epsilon_1 T$ . Applying the results of Jin et al. (2018), their  $\text{Regret}_{MDP}$  is  $\mathcal{O}(\sqrt{H^3 SAT}) = \mathcal{O}(\sqrt{\frac{1}{\epsilon_1} \cdot \frac{1}{\epsilon_1} T})$ . To minimize  $\text{Regret}_{total}$ , we balance the  $\text{Regret}_{MDP}$  and  $\text{Regret}_{gap}$  by setting  $\sqrt{\frac{1}{\epsilon_1} \cdot \frac{1}{\epsilon_1} T} = \epsilon_1 T$ , which gives  $\epsilon_1 = \frac{1}{T^{1/4}}$ , giving us an optimized regret bound of  $\mathcal{O}(T^{\frac{3}{4}} \sqrt{H^3 \log T})$ .

For Dong et al. (2019), suppose we discretize the state and action space optimally with step-size  $\epsilon_2$  to apply Dong et al. (2019) to the backlogged/lost-sales episodic inventory control problem. We also optimize aggregation using the special property of these inventory control problems that the Q-values only depend on the action not the state, so we aggregate all the state-action pairs  $(x_1, y), (x_2, y)$  into one aggregated state-action pair. This 0-error aggregation helps reduce the aggregated state-action space. Then the optimized regret bound in Dong et al. (2019) is  $\mathcal{O}(\sqrt{H^4 \frac{1}{\epsilon} T \log T} + \epsilon T)$ . We minimize  $\text{Regret}_{total}$  by balancing the two terms and take  $\epsilon = \frac{1}{T^{1/3}}$ , obtaining an optimized regret bound of  $\mathcal{O}(T^{\frac{2}{3}} \sqrt{H^4 \log T})$ .

## E MISSING PROOFS FOR FQL

For the proof for FQL, we adopt similar notations and flow of the proof in Jin et al. (2018) (but adapted to our full-feedback setting) to facilitate quick comprehension for readers who are familiar with Jin et al. (2018).

Like Jin et al. (2018), we use  $[\mathbb{P}_h V_{h+1}](x, y) := \mathbb{E}_{x' \sim \mathbb{P}(\cdot|x, y)} V_{h+1}(x')$ . Then the Bellman optimality equation becomes  $Q_h^*(x, y) = (r_h + \mathbb{P}_h V_{h+1}^*)(x, y)$ .

Similar to Equation 4 but without “skipping”,  $FQL$  updates the  $Q$  values in the following way for any  $(x, y) \in \mathcal{A}$  at any time step:

$$Q_h^{k+1}(x, y) \leftarrow (1 - \alpha_k) Q_h^k(x, y) + \alpha_k [r_h^{k+1}(x, y) + V_{h+1}^k(x_{h+1})] \quad (15)$$

Then by the definition of weights  $\alpha_t^k$ , we have

$$Q_h^k(x, y) = \alpha_{k-1}^0 H + \sum_{j=1}^{k-1} \alpha_{k-1}^j \left[ r_h^j(x, y) + V_{h+1}^j(x_{h+1}^j) \right] \quad (16)$$

The following two lemmas are variations of Lemma 5 and Lemma 6.

**Lemma 12.** For any  $(x, y, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$ , we have

$$\begin{aligned} (Q_h^k - Q_h^*)(x, y) &= \alpha_{k-1}^0 (H - Q_h^*(x, y)) \\ &\quad + \sum_{i=1}^{k-1} \alpha_{k-1}^i \left[ (V_{h+1}^i - V_{h+1}^*)(x_{h+1}^i) + r_h^i - \mathbb{E}[r_h^i] + \left[ (\hat{\mathbb{P}}_h^i - \mathbb{P}_h) V_{h+1}^* \right](x, y) \right] \end{aligned}$$

*Proof.* From the Bellman optimality equation  $Q_h^*(x, y) = \mathbb{E}[r_h(x, y)] + \mathbb{P}_h V_{h+1}^*(x, y)$ , our notation  $[\hat{\mathbb{P}}_h^i V_{h+1}^*](x, y) := V_{h+1}^*(x_{h+1}^i)$ , and the fact that  $\sum_{i=0}^{k-1} \alpha_{k-1}^i = 1$ , we have

$$Q_h^*(x, y) = \alpha_{k-1}^0 Q_h^*(x, y) + \sum_{i=1}^{k-1} \alpha_{k-1}^i \left[ \mathbb{E}[r_h^i(x, y)] + (\mathbb{P}_h - \hat{\mathbb{P}}_h^i) V_{h+1}^*(x, y) + V_{h+1}^*(x_{h+1}^i) \right]$$

Subtracting Equation 16 from this equation gives us Lemma 12.  $\square$

**Lemma 13.** For any  $p \in (0, 1)$ , with probability at least  $1 - p$ , for any  $(x, y, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$ , let  $\iota = \log(SAT/p)$ , we have for some absolute constant  $c$ :

$$0 \leq (Q_h^k - Q_h^*)(x, y) \leq \alpha_{k-1}^0 H + \sum_{i=1}^{k-1} \alpha_{k-1}^i (V_{h+1}^i - V_{h+1}^*)(x_{h+1}^i) + c \sqrt{\frac{H^3 \iota}{k-1}} \quad (17)$$

*Proof.* For any  $i \in [k]$ , recall that episode  $i$  is the episode where the state-action pair  $(x, y)$  was updated at stage  $h$  for the  $i$ th time. Let  $\mathcal{F}_h^i$  be the  $\sigma$ -field generated by all the random variables until episode  $i$ , stage  $h$ . Then for any  $\tau \in [K]$ ,  $\left( [\hat{\mathbb{P}}_h^i - \mathbb{P}_h] V_{h+1}^*(x, y) + r_h^i - \mathbb{E}[r_h^i] \right)_{i=1}^\tau$  is a martingale difference sequence w.r.t. the filtration  $\{\mathcal{F}_h^i\}_{i \geq 0}$ . Then by Azuma-Hoeffding Theorem, we have that with probability at least  $1 - p/SAT$ :

$$\left| \sum_{i=1}^{k-1} \alpha_{k-1}^i \cdot \left[ (\hat{\mathbb{P}}_h^i - \mathbb{P}_h) V_{h+1}^* \right](x, y) + r_h^i - \mathbb{E}[r_h^i] \right| \leq \frac{cH}{2} \sqrt{\sum_{i=1}^{k-1} (\alpha_{k-1}^i)^2 \cdot \iota} \leq c \sqrt{\frac{H^3 \iota}{k-1}} \quad (18)$$

for some constant  $c$ .

Now we union bound over states, actions and times, we see that with probability at least  $1 - p$ , we have

$$\left| \sum_{i=1}^{k-1} \alpha_{k-1}^i \left[ (\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h) V_{h+1}^* \right](x, y) + r_h^i - \mathbb{E}[r_h^i] \right| \leq c \sqrt{\frac{H^3 \iota}{k-1}} \quad (19)$$

Then the right-hand side of Lemma 13 follows from Lemma 12 and Inequality 19. The left-hand side also follows from Lemma 12 and Inequality 19 using induction on  $h = H, H-1, \dots, 1$ .  $\square$

*Proof.* (**Theorem ??**) Define  $\Delta_h^k := (V_h^k - V_h^{\pi_k})(x_h^k)$  and  $\phi_h^k := (V_h^k - V_h^*)(x_h^k)$ .

By Lemma 18, with  $1 - p$  probability,  $Q_h^k \geq Q_h^*$  and thus  $V_h^k \geq V_h^*$ . Thus the total regret can be upper bounded:

$$\text{Regret}(K) = \sum_{k=1}^K (V_1^* - V_1^{\pi_k})(x_1^k) \leq \sum_{k=1}^K (V_1^k - V_1^{\pi_k})(x_1^k) = \sum_{k=1}^K \Delta_1^k$$

The main idea of the rest of the proof is to upper bound  $\sum_{k=1}^K \Delta_h^k$  by the next step  $\sum_{k=1}^K \Delta_{h+1}^k$ , which gives a recursive formula to obtain the total regret. Let  $y_h^k$  denote the base stock levels taken at stage  $h$  of episode  $k$ , which means  $y_h^k = \arg \max Q_h^k(y')$ .

$$\begin{aligned} \Delta_h^k &= (V_h^k - V_h^{\pi_k})(x_h^k) \stackrel{(1)}{\leq} (Q_h^k - Q_h^{\pi_k})(x_h^k, y_h^k) \\ &= (Q_h^k - Q_h^*)(x_h^k, y_h^k) + (Q_h^* - Q_h^{\pi_k})(x_h^k, y_h^k) \\ &\stackrel{(2)}{\leq} \alpha_{k-1}^0 H + \sum_{i=1}^{k-1} \alpha_{k-1}^i \phi_{h+1}^i + c\sqrt{\frac{H^3 \ell}{k-1}} + [\mathbb{P}_h(V_{h+1}^* - V_{h+1}^{\pi_k})](x_h^k, y_h^k) \\ &= \alpha_{k-1}^0 H + \sum_{i=1}^{k-1} \alpha_{k-1}^i \phi_{h+1}^i + c\sqrt{\frac{H^3 \ell}{k-1}} + [(\mathbb{P}_h - \hat{\mathbb{P}}_h^k)(V_{h+1}^* - V_{h+1}^{\pi_k})](x_h^k, y_h^k) \\ &\quad + (V_{h+1}^* - V_{h+1}^{\pi_k})(x_{h+1}^k) \\ &\stackrel{(3)}{=} \alpha_{k-1}^0 H + \sum_{i=1}^{k-1} \alpha_{k-1}^i \phi_{h+1}^i + c\sqrt{\frac{H^3 \ell}{k-1}} - \phi_{h+1}^k + \Delta_{h+1}^k + \xi_{h+1}^k \end{aligned} \quad (20)$$

where  $\xi_{h+1}^k := [(\mathbb{P}_h - \hat{\mathbb{P}}_h^k)(V_{h+1}^* - V_{h+1}^{\pi_k})](x_h^k, y_h^k)$  is a martingale difference sequence. Inequality (1) holds because  $V_h^k(x_h^k) \leq \max_{\text{feasible } y'} Q_h^k(x_h^k, y') = Q_h^k(x_h^k, y_h^k)$ , and Inequality (2) holds by Lemma 13 and the Bellman equations. Inequality (3) holds by definition  $\Delta_{h+1}^k - \phi_{h+1}^k = (V_{h+1}^* - V_{h+1}^{\pi_k})(x_{h+1}^k)$ .

In order to compute  $\sum_{k=1}^K \Delta_1^k$ , we need to first bound the first term in Equation 20. Since  $\alpha_k^0 = 0, \forall k \geq 1$ , we know that  $\sum_{k=1}^K \alpha_{k-1}^0 H \leq H$ .

Now we bound the sum of the second term in Equation 7 over the episodes by regrouping:

$$\sum_{k=2}^K \sum_{i=1}^{k-1} \alpha_{k-1}^i \phi_{h+1}^i \leq \sum_{i=1}^{K-1} \phi_{h+1}^i \sum_{k=i+1}^{\infty} \alpha_{k-1}^i \leq \sum_{i=1}^{K-1} \phi_{h+1}^i \sum_{k'=i}^{\infty} \alpha_{k'}^i \leq \left(1 + \frac{1}{H}\right) \sum_{k=1}^K \phi_{h+1}^k \quad (21)$$

where the last inequality uses  $\sum_{t=i}^{\infty} \alpha_t^i = 1 + \frac{1}{H}$  for every  $i \geq 1$  from Lemma 3.

Plugging the above Equation 21 and  $\sum_{k=1}^K \alpha_k^0 H \leq H$  back into Equation 7, we have:

$$\begin{aligned} \sum_{k=1}^K \Delta_h^k &\leq H + \sum_{k=2}^K \Delta_h^k \\ &\leq H + H + \left(1 + \frac{1}{H}\right) \sum_{k=1}^K \phi_{h+1}^k - \sum_{k=2}^K \phi_{h+1}^k + \sum_{k=2}^K \Delta_{h+1}^k + \sum_{k=2}^K c\sqrt{\frac{H^3 \ell}{k-1}} + \sum_{k=2}^K \xi_{h+1}^k \\ &\leq 2H + \phi_{h+1}^1 + \frac{1}{H} \sum_{k=2}^K \phi_{h+1}^k + \sum_{k=2}^K \Delta_{h+1}^k + \sum_{k=2}^K c\sqrt{\frac{H^3 \ell}{k-1}} + \sum_{k=2}^K \xi_{h+1}^k \\ &\leq 3H + \left(1 + \frac{1}{H}\right) \sum_{k=2}^K \Delta_{h+1}^k + \sum_{k=2}^K c\sqrt{\frac{H^3 \ell}{k-1}} + \sum_{k=2}^K \xi_{h+1}^k \end{aligned} \quad (22)$$

where the last inequality uses  $\phi_{h+1}^k \leq \Delta_{h+1}^k$ . By recursing on  $h = 1, 2, \dots, H$ , and because  $\Delta_{H+1}^K = 0$ :

$$\sum_{k=1}^K \Delta_1^k \leq \mathcal{O} \left( \sum_{h=1}^H \sum_{k=1}^K (c \sqrt{\frac{H^3 \iota}{k-1}} + \xi_{h+1}^k) \right)$$

where  $\sum_{h=1}^H \sum_{k=1}^K c \sqrt{\frac{H^3 \iota}{k-1}} = \mathcal{O}(H \sqrt{H^3 \log(SAT/p)} \sqrt{K}) = \tilde{\mathcal{O}}(\sqrt{H^4 T})$ .

On the other hand, by Azuma-Hoeffding inequality, with probability  $1 - p$ , we have

$$\left| \sum_{h=1}^H \sum_{k=1}^K \xi_{h+1}^k \right| = \left| \sum_{h=1}^H \sum_{k=1}^K \left[ \left( \mathbb{P}_h - \hat{\mathbb{P}}_h^k \right) (V_{h+1}^* - V_{h+1}^{\pi_k}) \right] (x_h^k, y_h^k) \right| \leq cH \sqrt{T_l} \leq \tilde{\mathcal{O}}(\sqrt{H^4 T}) \quad (23)$$

which establishes  $\sum_{k=1}^K \Delta_1^k \leq \tilde{\mathcal{O}}(H^2 \sqrt{T})$ .  $\square$

## F MORE NUMERICAL EXPERIMENTS

We show more numerical experiment results to demonstrate the performance of *FQL* and *HQL*. In Table 3, we use again the backlogged model to compare *FQL* and *HQL* against *OPT*, *Aggregated QL* and *QL-UCB*, but with a different set of parameter than in Section 7. In Table 4 and 5, we use the lost-sales model to compare *HQL* against *OPT*, *Aggregated QL* and *QL-UCB*.

For Tables 3 and 4, we make the demand distribution less adversarial: with each step in the episode, we have demands that are increasing in expectation. However, we let the upper bound of base-stock levels increase with the episode length  $H$ , which is more adversarial. For Table 5, we use the same demand distributions and base-stock upper bound as in Table 2 in Section 7.

We run each experimental point 300 times for statistical significance.

**Episode length:**  $H = 1, 3, 5$ .

**Number of episodes:**  $K = 100, 500, 2000$ .

**Demands:**  $D_h \sim U[0, 1] + h$ .

**Holding cost:**  $o_h = 2$ .

**Backlogging cost:**  $b_h = 10$ .

**Action space:**  $[0, \frac{1}{20}, \frac{2}{20}, \dots, 2H]$ .

Table 3: Comparison of cumulative costs for backlogged episodic inventory control with less adversarial demands and increasing base-stock upper bounds

H	K	OPT		FQL		HQL		Aggregated QL		QL-UCB	
		mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
1	100	89.1	3.8	97.1	5.5	117.3	16.8	160.1	8.3	327.5	18.8
	500	420.2	4.2	431.2	4.2	507.8	45.6	732.7	22.1	825.4	10.9
	2000	1669.8	4.8	1691.2	6.6	1883.6	99.7	2546.2	32.6	2952.1	19.9
3	100	253.0	6.6	304.6	9.6	423.8	15.4	510.9	14.4	1712.0	19.1
	500	1252.4	7.0	1314.3	11.9	1611.0	43.9	1703.2	16.1	4603.7	101.6
	2000	5056.2	6.5	5128.7	10.2	5702.8	104.7	6188.0	14.1	15088.6	132.0
5	100	415.9	6.4	543.6	11.0	762.4	30.0	3011.8	1294.6	6101.9	357.6
	500	2077.1	12.7	2224.6	15.6	2746.3	113.7	10277.1	6888.5	11763.6	2982.5
	2000	8394.3	6.2	8557.2	11.1	9630.4	356.6	30489.8	31232.4	39873.8	7210.1

Again for *Aggregated QL* from Dong et al. (2019) and for *QL-UCB* from Jin et al. (2018), we optimize by taking the Q-values to be only dependent on the action, thus reducing the state-action pair space. As in Section 7, we do not fine-tune the confidence interval for *HQL* for different settings, but use a general formula  $\sqrt{\frac{H \log(HKA)}{k}}$  as the confidence interval for all settings. We also do not fine-tune the *UCB bonus* defined in *QL-UCB* (see Jin et al. (2018)).

A caveat of *Aggregated QL* from Dong et al. (2019) is that we need to know a good aggregation of the state-action pairs beforehand, which is usually unavailable for online problems. For using *Aggregated QL* in Table 3 and 4, we further aggregate the state and actions to be multiples of  $1/2$ . For using *Aggregated QL* in Table 5 (and also in Section 7), we further aggregate the state and actions to be multiples of 1.



**Episode length:**  $H = 1, 3, 5$ .**Number of episodes:**  $K = 100, 500, 2000$ .**Demands:**  $D_h \sim U[0, 1] + h$ .**Holding cost:**  $o_h = 2$ .**Lost-Sales Penalty:**  $b_h = 10$ .**Action space:**  $[0, \frac{1}{20}, \frac{2}{20}, \dots, 2H]$ .

Table 4: Comparison of cumulative costs for lost-sales episodic inventory control with less adversarial demands and increasing base-stock upper bounds

H	K	OPT		HQL		Aggregated QL		QL-UCB	
		mean	SD	mean	SD	mean	SD	mean	SD
1	100	89.1	3.8	117.3	16.8	201.7	6.6	291.7	6.6
	500	420.2	4.2	507.8	44.6	1002.8	4.0	1452.8	4.0
	2000	1669.8	4.8	1883.6	99.7	4012.1	5.3	5812.1	5.3
3	100	253.0	6.6	443.8	65.9	1902.8	81.4	2071.4	29.9
	500	1252.4	7.0	1730.7	361.3	9534.0	379.7	10375.7	13.1
	2000	5056.2	6.5	6163.2	374.3	38139.6	1519.4	41504.9	22.6
5	100	415.9	6.4	780.6	64.3	5716.6	153.0	5902.8	44.8
	500	2077.1	12.7	2926.0	332.6	28510.7	764.3	29385.1	183.1
	2000	8394.3	6.2	10560.1	1201.6	114010.7	3080.2	117481.6	727.6

**Episode length:**  $H = 1, 3, 5$ .**Number of episodes:**  $K = 100, 500, 2000$ .**Demands:**  $D_h \sim (10 - h)/2 + U[0, 1]$ .**Holding cost:**  $o_h = 2$ . **Lost-Sales Penalty:**  $b_h = 10$ . **Action Space:**  $[0, \frac{1}{20}, \frac{2}{20}, \dots, 10]$ .

Table 5: Comparison of cumulative costs for lost-sales episodic inventory control with the original demands and base-stock upper bounds

H	K	OPT		HQL		Aggregated QL		QL-UCB	
		mean	SD	mean	SD	mean	SD	mean	SD
1	100	88.2	4.1	125.9	19.2	705.4	9.7	895.4	9.7
	500	437	4.4	528.9	44.1	3506.1	4.4	4456.1	4.4
	2000	1688.9	2.8	1929.2	89.1	14005.6	6.6	17805.6	6.6
3	100	257.4	3.2	448.4	52.1	2405.6	9.1	2975.6	9.1
	500	1274.6	6.1	1746.7	239.9	12009.3	6.4	14859.3	6.4
	2000	4965.6	8.3	6111.2	918.2	47926.4	14.8	59326.4	14.8
5	100	421.2	3.3	774.6	51.8	4497.4	11.6	5447.4	11.6
	500	2079.0	8.2	2973.9	299.9	22478.5	10.7	27228.5	10.7
	2000	8285.7	8.3	10701.1	1207.5	89929.7	14.0	108929.7	14.0

As we can see in all of our experiments, *FQL* and *HQL* both perform very promisingly with significant advantage over the other two existing algorithms. *FQL* stays consistently very close to the clairvoyant optimal in both the more adversarial and less adversarial settings for the backlogged model. *HQL* catches up rather quickly to *OPT* in all the settings for both the backlogged model and the lost-sales model.