Appendix A: Regret Analysis for Elimination-Based Half-Q-Learning

Recall the coefficients $\alpha_t := \frac{H+1}{H+t}$ used in Algorithm *HQL*. We define related weights $\alpha_t^0 := \prod_{j=1}^t (1-\alpha_j)$, and $\alpha_t^i := \alpha_i \prod_{j=i+1}^t (1-\alpha_j)$ as in Jin et al. (2018) and in Dong et al. (2019). Below are useful properties of these weights:

LEMMA 5. *The following properties hold:*

1. $\sum_{i=1}^t \alpha_t^i = 1$ *and* $\alpha_t^0 = 0, \forall t \geq 1$*;*
2. $\sum_{i=1}^t \alpha_t^i = 0$ *and* $\alpha_t^0 = 1$ *when* $t = 0$*;*
3. $\max_{i\in[t]} \alpha_t^i \leq \frac{2H}{t}$ *and* $\sum_{i=1}^t (\alpha_t^i)^2 \leq \frac{2H}{t}, \forall t \geq 1$*;*
4. $\sum_{t=i}^\infty \alpha_t^i = 1 + \frac{1}{H}$ *for every* $i \geq 1$*;*
5. $\frac{1}{\sqrt{t}} \leq \sum_{i=1}^t \frac{\alpha_t^i}{\sqrt{i}} \leq \frac{1+\frac{1}{H}}{\sqrt{t}}$ *for every* $t \geq 1$*.*

Remark: The last property is tighter than the corresponding bound in Lemma 4.1 in Jin et al. (2018). See Appendix A.1.

We state the fact that base-stock policies are optimal for the episodic lost-sales model with zero lead time in the following Lemma 6. Lemma 6 can be obtained by applying classical results in Porteus (2002). For completeness, we provide a proof in Appendix A.2.

LEMMA 6. *Base-stock policies are optimal for the episodic lost-sales model with zero lead time.*

For any base-stock policy, the reward and the leftover inventory level only depend on the base-stock level and do not depend on the state, even though the feasible action set depends on the state. Therefore, in this setting, we can simplify the Q-value functions: $Q(x,y) = Q(y), \forall x \in \mathcal{S}$.

Recall for any $(x,h,k) \in \mathcal{S} \times [H] \times [K]$, and for any base-stock level $y \in A_h^k$, $\tau_h^k(x,y)$ is the next time step after time step h in episode k that our policy lands on a simulated inventory level $x'^k_{\tau_h^k(x,y)}$ that allows us to take an action in the running set $A^k_{\tau_h^k(x,y)}$. Therefore, $\tau_h^k(x,y)$ is a stopping time. The time steps in between are "skipped" in the sense that the Q,V-values for those time steps never appear on the right hand side of Equation (8) when we update value functions. If no skipping happened, then $\tau_h^k(x,y) = h+1$, and we have the original Bellman equation (2.1). Using the general property of optional stopping that $\mathbf{E}[M_\tau] = M_0$ for any stopping time τ and discrete-time martingale M_τ, our Bellman optimality equation becomes the following *delayed form* of the Bellman equation:

$$Q_h^*(x,y) = Q_h^*(y) = \mathbf{E}_{\tau_h^k, \tilde{r}^*_{h,\tau_h^k}, x'_{\tau_h^k}} [\tilde{r}^*_{h,\tau_h^k} + V^*_{\tau_h^k}(x'_{\tau_h^k})] \tag{7}$$

where we simplify the notation $\tau_h^k(x,y)$ to τ_h^k, and recall $\tilde{r}_{h,h'}$ denotes the cumulative reward from step h to h'.

Using the stopping times and simulated trajectories, *HQL* updates the Q-values backward $h = H, \ldots, 1$ as follows:

$$Q_h^{k+1}(y) \leftarrow (1-\alpha_k) Q_h^k(y) + \alpha_k [\tilde{r}^{k+1}_{h,\tau_h^{k+1}} + V^{k+1}_{\tau_h^{k+1}}(x'_{\tau_h^{k+1}})]. \tag{8}$$

where Q_h^k, V_h^k denotes the Q_h, V_h functions at the beginning of episode k respectively.

Then by Equation (8) and the definition of the weights α_k^i's,

$$Q_h^k(y) = \alpha_{k-1}^0 H + \sum_{i=1}^{k-1} \alpha_{k-1}^i \left[\tilde{r}_{h,\tau_h^k}^i + V_{\tau_h^k}^{i+1}\left(x_{\tau_h^k}^i\right)\right]. \tag{9}$$

which naturally gives us Lemma 7, where we bound the difference between the optimal Q-value of a state-action pair and our estimated Q-value. The proof of Lemma 7 is provided in Appendix A.3.

LEMMA 7. *For any $(x,h,k) \in \mathcal{S}\times[H]\times[K]$, and for any $y \in A_h^k$, we have*

$$\begin{aligned}\left(Q_h^k - Q_h^*\right)(y) =& \alpha_{k-1}^0 \left(H - Q_h^\star(y)\right) + \sum_{i=1}^{k-1} \alpha_{k-1}^i \Big[\left(V_{\tau_h^i}^{i+1} - V_{\tau_h^i}^*\right)(x_{\tau_h^i}^i) + \tilde{r}_{h,\tau_h^i}^i - \tilde{r}_{h,\tau_h^i}^* \\ &+ \left(V_{\tau_h^i(x,y)}^*(x_{\tau_h^i}^i) + \tilde{r}_{h,\tau_h^i}^* - \mathbf{E}_{\tilde{r}^*,x',\tau_h^i}\left[\tilde{r}_{h,\tau_h^i}^* + V_{\tau_h^i}^*(x'_{\tau_h^i})\right]\right)\Big].\end{aligned}$$

Then by identifying the martingales in the right-hand side of Lemma 7, we bound the difference between our Q-value estimates and the optimal Q-values in the following lemma:

LEMMA 8. *For any $(x,h,k) \in \mathcal{S}\times[H]\times[K]$, and any $y \in A_h^k$, let $\iota = 9\log(AT)$, we have:*

$$\left|(Q_h^k - Q_h^*)(y)\right| \le \alpha_{k-1}^0 H + \sum_{i=1}^{k-1} \alpha_{k-1}^i \left|\left(V_{\tau_h^i}^{i+1} - V_{\tau_h^i}^*\right)(x_{\tau_h^i}^i) + \tilde{r}_{h,\tau_h^i}^i - \tilde{r}_{h,\tau_h^i}^*\right| + c\sqrt{\frac{H^3\iota}{k-1}} \tag{10}$$

with probability at least $1 - 1/(AT)^8$, for some $c \ge 2\sqrt{2}$.

The proof of Lemma 8 is provided in Appendix A.4.

We review *shortfall decomposition* below. For proof and reference, see Appendix A.7.

LEMMA 9. *(shortfall decomposition) For any policy π and any episode k, the per-episode regret is:*

$$\left(V_1^* - V_1^{\pi_k}\right)(x_1^k) = \mathbf{E}_\pi\big[\sum_{h=1}^H \big(\max_{y\in\mathcal{A}} Q_h^*(x_h^k, y) - Q_h^*(x_h^k, y_h^k)\big)\big].$$

Shortfall decomposition allows us to calculate the regret of our policy by summing up the difference between the optimal Q-values of our action and those of the optimal action from the same state. We then find high-probability upper-bounds on the sum.

Proof for Theorem 1 Recall that we partition the time steps $h = 1,\dots,H$ in each episode k into two sets, Γ_A^k and Γ_B^k, where Γ_A^k contains all the steps h where we are able to choose from the running set, and Γ_B^k contains all the steps h where we are unable to choose from the running set.

Then by shortfall decomposition, we have that the per-episode regret is

$$\begin{aligned}\left(V_1^* - V_1^{\pi_k}\right)(x_1^k) =& \mathbf{E}\Big[\sum_{h=1}^H \Big(\max_{y\ge x_h^k} Q_h^*(y) - Q_h^*(y_h^k)\Big)\Big] \\ \le& \mathbf{E}\Big[\sum_{h\in\Gamma_A^k} \max_{y\ge x_h^k}\Big(Q_h^*(y) - Q_h^*(y_h^k)\Big)\Big] + \mathbf{E}\Big[\sum_{h\in\Gamma_B^k} \max_{y\ge x_h^k}\Big(Q_h^*(y) - Q_h^*(y_h^k)\Big)\Big].\end{aligned}$$

Recall Lemma 1 and that we define $\{\delta_h\}_{h=1}^{H+1}$ to be a list of values that satisfy the following recursive relationship:

$$\delta_h = H + (1+1/H)\delta_{h+1} + c\sqrt{H^3\iota}, \forall h \in [H],$$

$$\delta_{H+1} = 0$$

where c is the same constant as in Lemma 8.

By Lemma 1.3, we can bound the first term on the right-hand side:

$$\begin{aligned}
&\mathbf{E}\Big[\sum_{h\in\Gamma_A^k}\max_{y\geq x_h^k}\Big(Q_h^*(y)-Q_h^*(y_h^k)\Big)\Big]\\
&\leq \mathbf{E}\Big[\sum_{h\in\Gamma_A^k}\frac{16\sqrt{H^5\iota}}{\sqrt{k-1}}\Big]\cdot\mathbf{P}\Big(\max_{y\geq x_h^k}\big(Q_h^*(y)-Q_h^*(y_h^k)\big)\leq\frac{16\sqrt{H^5\iota}}{\sqrt{k-1}}\Big)+\sum_{h\in\Gamma_A^k}H\cdot\mathbf{P}\Big(\max_{y\geq x_h^k}\Big(Q_h^*(y)-Q_h^*(y_h^k)\Big)>\frac{16\sqrt{H^5\iota}}{\sqrt{k-1}}\Big)\\
&\leq \mathcal{O}\Big(\sum_{h\in\Gamma_A^k}\frac{\sqrt{H^5\iota}}{\sqrt{k-1}}\Big)+\mathcal{O}\Big(\sum_{h\in\Gamma_A^k}\frac{H}{A^5T^5}\Big).
\end{aligned}$$

By Lemma 1.4, we can bound the last term

$$\mathbf{E}\Big[\sum_{h\in\Gamma_B^k}\max_{y\geq x_h^k}\Big(Q_h^*(y)-Q_h^*(y_h^k)\Big)\Big]\leq 0\cdot\Big(1-\frac{1}{A^5T^5}\Big)+\sum_{h\in\Gamma_B^k}H\cdot\frac{1}{A^5T^5}\leq\sum_{h\in\Gamma_B^k}H\cdot\frac{1}{A^5T^5}.$$

Then the difference between the expected total reward of *HQL* and of the optimal policy π^* is

$$\begin{aligned}
\mathrm{Regret}_{MDP}(K)&=\left(V_1^*-V_1^{\pi_1}\right)(x_1^1)+\sum_{k=2}^{K}\left(V_1^*-V_1^{\pi_k}\right)(x_1^k)\\
&\leq H+\sum_{k=2}^{K}\mathcal{O}\Big(\sum_{h\in\Gamma_B^k}\frac{H}{A^5T^5}+\sum_{h\in\Gamma_A^k}\frac{\sqrt{H^5\iota}}{\sqrt{k-1}}+\sum_{h\in\Gamma_A^k}\frac{H}{A^5T^5}\Big)\leq\sum_{k=2}^{K}\frac{\mathcal{O}(\sqrt{H^7\iota})}{\sqrt{k-1}}\leq\mathcal{O}(H^3\sqrt{T\iota}).
\end{aligned}$$

It follows that the total expected regret of *HQL* against *OPT* is

$$\mathrm{Regret}_{total}(K)=\mathrm{Regret}_{MDP}(K)+\mathrm{Regret}_{gap}(K)=\mathcal{O}\Big(H^3\sqrt{T\iota}+(M-m)/K\Big)=\mathcal{O}\Big(H^3\sqrt{T\log T}\Big)$$

Finally, recall the scaling we performed on the reward, so we multiply by the factor $\mathcal{O}\big(M\cdot\max(|o_h|,|p_h|)\big)$. This implies an $\mathcal{O}\big(H^3M\cdot\max(|o_h|,|p_h|)\sqrt{T\log T}\big)$ total dependence on all setting parameters.

A.1. Properties of weights α_t^i

We obtain the last property in Lemma 5 by a more careful algebraic analysis, so that we obtain a tighter bound on $\sum_{i=1}^t\frac{\alpha_t^i}{\sqrt{i}}$ than the corresponding bound in Jin et al. (2018). For the remaining properties in Lemma 5, see Lemma 4.1 in Jin et al. (2018).

Proof of Lemma 5, part 5 We prove the last property in Lemma 5 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 $i=1,\ldots,t-1$ we have

$$\sum_{i=1}^{t}\frac{\alpha_t^i}{\sqrt{i}}=\frac{\alpha_t}{\sqrt{t}}+(1-\alpha_t)\sum_{i=1}^{t-1}\frac{\alpha_{t-1}^i}{\sqrt{i}}\tag{11}$$

Assuming the inductive hypothesis holds, on the one hand,

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

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

$$\begin{aligned}
\frac{\alpha_t}{\sqrt{t}}+(1-\alpha_t)\sum_{i=1}^{t-1}\frac{\alpha_{t-1}^i}{\sqrt{i}}&\leq\frac{\alpha_t}{\sqrt{t}}+\frac{(1+1/H)(1-\alpha_t)}{\sqrt{t-1}}=\frac{H+1}{\sqrt{t}(H+t)}+\frac{(1+1/H)\sqrt{t-1}}{H+t}\\
&\leq\frac{H+1}{\sqrt{t}(H+t)}+\frac{(1+1/H)\sqrt{t}}{H+t}\leq\frac{(1+1/H)}{\sqrt{t}}
\end{aligned}\tag{12}$$

where the first inequality holds by the inductive hypothesis. □

A.2. Optimality of base-stock policies

Proof of Lemma 6 Since the optimal value functions $V_h^*(\cdot)$ and $Q_h^*(\cdot)$ evaluate all possible ways of ordering inventory at each time step throughout each episode, the fact that they turn out to be concave (Lemma 10) implies that there is one single quantity that we should order up to for each time period h to obtain the maximum expected reward. □

A.3. Proof of Lemma 7

Proof of Lemma 7 From the Bellman optimality equation (7), and the fact that $\sum_{i=0}^{k-1}\alpha_{k-1}^i = 1$, we have

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

Subtracting Equation (9) from this equation, and adding some of the middle terms that cancel with themselves gives us Lemma 7. □

A.4. Proof of Lemma 8

Proof of Lemma 8 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 σ-field generated by all the random variables until episode i, stage h. Then for any $\tau \in [K]$,

$$\left(V_{\tau_h^i}^*(x_{\tau_h^i}^i) + \tilde{r}_{\tau_h^i}^* - \mathbb{E}_{\tilde{r}^*,x',\tau_h^i}\left[\tilde{r}_{\tau_h^i}^* + V_{\tau_h^i}^*(x'_{\tau_h^i})\right]\right)_{i=1}^{\tau}$$

is a martingale difference sequence with respect to the filtration $\{\mathcal{F}_h^i\}_{i\geq 0}$. Then by Azuma-Hoeffding Theorem, we have that with probability at least $1-(1/AT)^9$:

$$\left|\sum_{i=1}^{k-1}\alpha_{k-1}^i\cdot\left(V_{\tau_h^i}^*(x_{\tau_h^i}^i) + \tilde{r}_{\tau_h^i}^* - \mathbb{E}_{\tilde{r}^*,x',\tau_h^i}\left[\tilde{r}_{\tau_h^i}^* + V_{\tau_h^i}^*(x'_{\tau_h^i})\right]\right)\right| \leq \frac{cH}{2}\sqrt{\sum_{i=1}^{k-1}\left(\alpha_{k-1}^i\right)^2\cdot\iota} \leq c\sqrt{\frac{H^3\iota}{k-1}} \tag{13}$$

for any constant $c \geq 2\sqrt{2}$. By union bound, we have with probability at least $1-(1/AT)^8$ that for any $x, h, k, y \in A_h^k$,

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

By this equation and Lemma 7, Lemma 8 follows. □

A.5. Upper bound on sequence $\delta_h, h=1,\ldots,H$ in Equation (5)

Proof: We set $d_h = (\delta_h)\cdot\left(1+\frac{1}{H}\right)^h$ and observe that the recurrence implies

$$d_h = d_{h+1} + H + 2\sqrt{2}\sqrt{H^3\iota} \tag{14}$$

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

A.6. Concavity of the Optimal Value Functions

Below we prove the concavity of the Q, V value functions of the lost-sales model with zero lead time. The same proof and result applies to the single-product backlogged model.

LEMMA 10. *For the lost-sales model, 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.*

Proof: We proceed by backward induction on h, starting from the base case $h = H$. The base case is the value functions for the last step of each episode: $Q_H^*(y)$ and $V_H^*(x)$. Since $Q_H^*(y)$ is just the expectation of a one time reward for the last period, we know $Q_H^*(y) = -[o_H(y - D_H)^+ + p_H \min(y, D_H)]$. This function is concave in y. Since $V_H^*(x) = \max_{y \geq x} Q_H^*(y)$, the graph of $V_H^*(x)$ is constant on the left of $x = \arg\max_{y \geq x} Q_H^*(y)$, and then goes down with a slope of $-o_H$ on the right of $x = \arg\max_{y \geq x} Q_H^*(y)$. So $V_H^*(x)$ is also concave.

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

Since $Q_h^*(y) = \mathbb{E}[V_{h+1}^*(y - D_h) + r_h(y, D_h)]$, and we know $r_h(y, D_h)$ is concave in y just like $Q_H^*(y)$, and that $V_{h+1}^*(x)$ is concave in x from the 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, as a weighted average of concave functions. Thus, $Q_h^*(y)$ is concave, and $V_h^*(x) = \max_{y \geq x} Q_h^*(y)$ is concave. $\square$

A.7. Shortfall decomposition

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

Proof of Lemma 9 For any policy π, let y_h^k denote the action the policy π_k takes at stage h of episode k. Let R_h denote the expected reward of y_h^k.

Define

$$Z_{h+1} = \begin{cases} R_h + \max_y Q_{h+1}^* \left(x_{h+1}^k, y\right) & \text{if } h < H \\ R_h & \text{if } h = H \end{cases}$$

Then

$$\mathbb{E}_\pi \left[Q_h^* \left(x_h^k, y_h^k\right)\right] = \mathbb{E}_\pi \left[Z_{h+1}\right]$$

Therefore,

$$\begin{aligned} V_1^* - V_1^{\pi_k} =& \mathbb{E}_\pi \Big[\max_{a \in \mathcal{A}} Q_1^*(x_1^k, a) - \sum_{h=1}^H R_h \Big] \\ =& \mathbb{E}_\pi \Big[\max_{a \in \mathcal{A}} Q_1^* \left(x_1^k, a\right) - \sum_{h=1}^H \left(R_h - Z_{h+1} + Q_h^* \left(x_h^k, y_h^k\right)\right) \Big] \\ =& \mathbb{E}_\pi \Big[\sum_{h=1}^H \left(\max_{a \in \mathcal{A}} Q_h^*(x_h^k, a) - Q_h^*(x_h, y_h^k)\right) \Big] \end{aligned}$$

$\square$

A.8. Proof for Lemma 1

Proof We prove by backward induction on $h = H+1, H, \ldots, 1$. 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 statements of the Lemma. Our proof goes by induction; for the base case $h = H+1$, we have $\delta_{H+1} = 0$ satisfies the Inequality in Lemma 1.1 (actually equality here) with probability 1 based on Bellman equations.

Now suppose that for some h, all parts of Lemma 1 hold for all $k \in [K]$ and $h' > h$. We will deduce from this that all parts hold for h as well. Note that since $\tau_h^k(x,a) > h$, the induction hypothesis implies that for all $a \in A_h^k$, with high probability:

$$\max_{y \in A^k_{\tau_h^k(x,a)}} \left| (Q^k_{\tau_h^k(x,a)} - Q^*_{\tau_h^k(x,a)})(y) \right| \le \frac{\delta_{\tau_h^k(x,a)}}{\sqrt{k-1}}. \tag{15}$$

We first show that Lemma 1.1 holds for h, which is the main step of the induction. By Lemma 8, with probability at least $1 - 1/(AT)^8$

$$\max_{y \in A_h^k} \left| (Q_h^k - Q_h^*)(y) \right| \le \max_{a \in A_h^k} \left\{ \alpha_{k-1}^0 H + \sum_{i=1}^{k-1} \alpha_{k-1}^i \left[\left(V^{i+1}_{\tau_h^i(x,a)} - V^*_{\tau_h^i(x,a)} \right) \left(x^i_{\tau_h^i(x,a)} \right)' + \tilde{r}^i_{h,\tau_h^i(x,a)} - \tilde{r}^*_{h,\tau_h^i(x,a)} \right] + c\sqrt{\frac{H^3\iota}{k-1}} \right\}$$

Based on our inductive hypothesis, we have

$$\max_{a \in A_h^k} \left[\left(V^{i+1}_{\tau_h^i(x,a)} - V^*_{\tau_h^i(x,a)} \right) \left(x^i_{\tau_h^i(x,a)} \right)' + \tilde{r}^i_{h,\tau_h^i(x,a)} - \tilde{r}^*_{h,\tau_h^i(x,a)} \right] \le \max_{y \in A^i_{\tau_h^i(x,a)}} |(Q^{i+1}_{\tau_h^i(x,a)} - Q^*_{\tau_h^i(x,a)})(y)| \le \frac{\delta_{\tau_h^i(x,a)}}{\sqrt{i}}$$

where the first inequality is because $\tilde{r}^i_{h,\tau_h^i(x,a)} - \tilde{r}^*_{h,\tau_h^i(x,a)}$ is with high probability zero because of the Lemma 1.4 part of the inductive hypothesis. Then

$$\max_{y \in A_h^k} \left| (Q_h^k - Q_h^*)(y) \right| \le \max_{a \in A_h^k} \left\{ \alpha_{k-1}^0 H + \left(\sum_{i=1}^{k-1} \alpha_{k-1}^i \cdot \frac{\delta_{\tau_h^i(x,a)}}{\sqrt{i}} \right) + c\sqrt{\frac{H^3\iota}{k-1}} \right\}. \tag{16}$$

We can bound α_{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 5:

$$\max_{y \in A_h^k} \left| (Q_h^k - Q_h^*)(y) \right| \le \frac{1}{\sqrt{k}} H + \frac{1+1/H}{\sqrt{k-1}} \delta_{\tau_h^i(x,a)} + c\sqrt{\frac{H^3\iota}{k}} \le \frac{1}{\sqrt{k-1}} H + \frac{1+1/H}{\sqrt{k-1}} \delta_{h+1} + c\sqrt{\frac{H^3\iota}{k-1}} = \frac{\delta_h}{\sqrt{k-1}} \tag{17}$$

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

Next we deduce that Lemma 1.2 holds for all $k \in [K]$ and the new value h. Recall for any (x,h,k) the definition $y_h^{k*} = \arg\max_{y \in A_h^k} Q_h^k(y)$. Supposing for sake of contradiction $y_h^* \notin A_h^k$, then we would have $Q_h^k(y_h^*) < Q_h^k(y_h^{k*}) - \frac{8\sqrt{H^5\iota}}{\sqrt{k-1}} = Q_h^k(x, y_h^{k*}) - \frac{2\delta_h}{\sqrt{k-1}}$. Hence either $Q_h^k(y_h^*) < Q_h^*(y_h^*) - \frac{\delta_h}{\sqrt{k-1}}$ or $Q_h^k(y_h^{k*}) > Q_h^*(y_h^{k*}) + \frac{\delta_h}{\sqrt{k-1}}$ must hold, violating Lemma 1.1. Therefore by Equation (15), $\text{Prob}(y_h^* \notin A_h^k(x)) \le \frac{1}{(AT)^5}$, i.e. the optimal action y_h^* is in the running set A_h^k with high probability.

Next we show Lemma 1.3 for the new value h. We will repeatedly use the bound (5). By the already-established Lemma 1.1 with new value h, we have with high probability:

$$Q_h^*(y_h^*) \le Q_h^k(y_h^{k*}) + \frac{\delta_h}{\sqrt{k-1}} \le Q_h^k(y_h^{k*}) + \frac{4\sqrt{H^5\iota}}{\sqrt{k-1}}.$$

Moreover recalling the upper confidence bound value, any action $y \in A_h^k$ satisfies

$$Q_h^k(y_h^{k*}) - Q_h^k(y) \le \frac{8\sqrt{H^5\iota}}{\sqrt{k-1}}.$$

Combining and using Lemma 1.1 again, we find (as desired) that

$$Q_h^*(y_h^*) \le \frac{12\sqrt{H^5\iota}}{\sqrt{k-1}} + \max_{y\in A_h^k(x)} Q_h^k(y) \le \frac{16\sqrt{H^5\iota}}{\sqrt{k-1}} + \max_{y\in A_h^k(x)} Q_h^*(y).$$

Finally we deduce Lemma 1.4 is true. From Lemma 1.2, we know that with high probability, the optimal action is in the running set. When the running set is not feasible to choose from, 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 with high probability. □

A.9. Regret caused by discretization.

Proof of Lemma 2 If we discretize $[m,M]$ with step-size $\frac{M-m}{T^2}$, for example, then $A=\Theta(T^2)$. Discretization incurs additional regret: $\text{Regret}_{gap}=\mathcal{O}(\frac{M-m}{T^2}\cdot HT)=o(1)$ by Lipschitzness of the reward function.

Appendix B: Applying existing Q-learning algorithms on the inventory control problems

Here we show that existing Q-learning results in general MDPs give suboptimal guarantees when specialized to our setting, as discussed in Section 3 of the main text.

For Jin et al. (2018), suppose we discretize the state and action space optimally with step-size ϵ_1 to apply Jin et al. (2018) to the backlogged/lost-sales episodic inventory control problem with continuous action and state space. Then the Regret_{gap} we get is $\epsilon_1 T$. Applying the results of Jin et al. (2018), their Regret_{MDP} is $\mathcal{O}(\sqrt{H^3SAT\iota})=\mathcal{O}(\sqrt{\frac{1}{\epsilon_1}\cdot\frac{1}{\epsilon_1}T\iota})$. To minimize Regret_{total}, we balance the Regret_{MDP} and 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 ϵ_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 Regret_{total} by balancing the two terms and take $\epsilon=\frac{1}{T^{1/3}}$, obtaining an optimized regret bound of $\mathcal{O}(T^{\frac{2}{3}}\sqrt{H^4\log T})$.

Appendix C: The Non-Discarding Lost-Sales Model

In the infinite-horizon version of the non-discarding lost-sales model with cyclic demands: to have finite V-values, a long-time average reward $\overline{r}$ is subtracted from the right-hand side of the Bellman equations:

$$V_t=\mathbb{E}[V_{t+1}+r_t]-\overline{r}.$$

(Puterman 2014, Theorem 8.4.7) guarantees the existence of an optimal average-reward policy. By taking limits of finite-horizon optimal policies, it can be proved that cyclic base-stock policies are optimal for the infinite-horizon problem.

Proof of Proposition 1 For inventory problems with known cyclic stochastic demands, (Zipkin 1989, Proposition 1c) shows the existence of a time $h \in [H]$ such that for the episodic problem with demand distributions $D_{h+1}, \ldots, D_H, D_1, \ldots, D_h$, the optimal base-stock level is maximized at the first round with demand D_{h+1} (referred to therein as the "maximal property"). For this choice of h, it readily follows that the base-stock levels for the episodic problem are equal to those of any repeated version of length $T = KH$ again started from D_{h+1}. Indeed, because the base-stock level B_{h+1} for the episodic problem is maximal, using the episodic base-stock policy repeatedly on the T-horizon problem is equivalent to solving K separate episodic problems - we are always able to order back up to exactly B_{h+1}. As a result, this algorithm solves each episode optimally while achieving a best-case initialization for each episode. This implies that it solves the T-horizon problem optimally for any $T = KH$. (However note that this T-horizon problem is shifted from the original.) □

We prove in the following proposition that the optimal policy for the infinite-horizon problem is also near optimal for the finite-horizon problem.

PROPOSITION 2. *For any h and sequence $(D_1, \ldots, D_h)$, the infinite-horizon optimal policy, denoted by π^*_∞, when applied to the finite-horizon problem, achieves expected regret $O(M\gamma)$ independent of the time horizon length T from any starting state x and time $h \in [H]$, with respect to the optimal finite-horizon policy, denoted by π^*_T.*

Proof: Suppose not, which means that the infinite-horizon optimal policy has some amount of regret C' that is larger than $O(M\gamma)$ after time T. We will construct a candidate infinite-horizon policy π' with superior performance to the optimal policy π^*_∞, which would be a contradiction to the definition.

We construct this candidate policy by the following 3 phases

1. Run the optimal T-horizon policy until time T.
2. Order nothing until all inventory is depleted.
3. Copy the infinite-horizon policy from the best possible starting point for the rest of time.

Since all states are reachable from a 0 inventory state, after all inventory is depleted by phase 2, all states are reachable in phase 3. Hence the above policy is feasible.

By assumption, phase 1 above achieves reward C' greater than the optimal policy on average. Meanwhile, phase 2 requires time $O(M\gamma)$ in expectation. Therefore, the candidate policy above eventually matches the trajectory of the infinite horizon policy, but its reward is larger by a positive constant $C - O(M\gamma) > 0$. Moreover, it has the same starting point. This is a contradiction because π^*_∞ is by definition the optimal policy for the infinite horizon problem. □

Proof of Lemma 3 To handle switches between arms, we simply wait for inventory to go below the base-stock level we want to choose for the beginning step of the next remaining arm h', and then start pulling arm h' once possible. By Assumption 3 for the non-episodic model, we know that each switch from an arm h to an arm h' will take $\mathcal{O}(M\gamma)$ time periods in expectation. By Markov Inequality, we know that the probability that the switch takes more than $\mathcal{O}(M\gamma)$ time periods is less than $1/2$. Then the probability that the switch takes more than $\mathcal{O}(M\gamma \cdot 3\log T)$ time periods is less than $(\frac{1}{2})^{3\log T} = \frac{1}{T^3}$.

Each phase j contains $\mathcal{O}(2^j H)$ time periods, so there are no more than $\log T$ phases. Since $|W_j| \le H$ for any j, there are only $\mathcal{O}(H \log T)$ arm switches in the whole horizon. Each switch takes time $\mathcal{O}(\gamma \log T)$ time periods with probability $1 - T^{-3}$, so switching between arms contribute negligible $\mathcal{O}(H\gamma \log^2 T)$ regret in the whole horizon. □

To analyze the regret bound for *Meta-HQL*, we need a tighter analysis than what is used in shortfall decomposition in *HQL*. Recall V_1^* denotes the expected optimal per-episode reward for the optimal policy. We use $V_1^{(w)}$ to denote the realized per-episode reward of the *w-shifted HQL* that arm w represents.

Proof of Lemma 4 First we want to show that for each arm, our estimated per-episode reward is very close to the true optimal per-episode reward for that arm. Let R_h denote the realized reward of y_h^k. For each episode $k \in [K]$ and time step $h \in [H]$, define

$$Z_{h+1}^k = \begin{cases} R_h^k + \max_y Q_{h+1}^* \left(x_{h+1}^k, y\right) & \text{if } h < H \\ R_h^k & \text{if } h = H \end{cases}$$

Then we have that for episode k, the difference between the optimal expected per-episode reward for the best arm and our realized per-episode reward for arm w is

$$\begin{aligned} V_1^* - V_1^{(w)} &= \max_a Q_1^*(x_1, a) - \sum_{h=1}^H R_h \\ &= \max_a Q_1^*(x_1, a) - \sum_{h=1}^H \left(R_h - Z_{h+1}^k + Q_h^*(x_h, y_h)\right) + \sum_{h=1}^H \left(Q_h^*(x_h, y_h) - Z_{h+1}\right) \end{aligned} \tag{18}$$

Consider the last term in Equation (18) for rounds $k = 1, \ldots, K_j$. Each of these terms is bounded between $[-H, H]$ and has mean 0 conditioned on the past. Therefore, the partial sums over $k = 1, \ldots, K_j$ are martingales, with the difference between consecutive martingales bounded by $[-H, H]$. For notation, we use superscript k to denote round k. By Azuma-Hoeffding Inequality,

$$\mathbb{P}\Big[\left|\sum_{k=1}^{K_j}\sum_{h=1}^{H}\left(Q_h^*(x_h^k, y_h^k) - Z_{h+1}^k\right)\right| \ge \epsilon\Big] \le 2\exp\left(-\frac{2\epsilon^2}{\sum_1^{K_j H} H^2}\right) \tag{19}$$

Then we have that the difference between the total expected reward and our realized reward for any arm if that arm is pulled for K_j rounds (meaning K_j cycles) is

$$\begin{aligned} \left|\sum_{k=1}^{K_j}(V_1^* - V_1^{\pi_k})\right| &\le \left|\sum_{k=1}^{K_j}\max_a Q_1^*(x_1^k, a) - \sum_{k=1}^{K_j}\sum_{h=1}^{H}\left(R_h - Z_{h+1}^k + Q_h^*(x_h^k, y_h^k)\right)\right| + \left|\sum_{k=1}^{K_j}\sum_{h=1}^{H}\left(Q_h^*(x_h^k, y_h^k) - Z_{h+1}^k\right)\right| \\ &\le \left|\sum_{k=1}^{K_j}\max_a Q_1^*(x_1^k, a) - \sum_{k=1}^{K_j}\sum_{h=1}^{H}\left(R_h - Z_{h+1}^k + Q_h^*(x_h^k, y_h^k)\right)\right| + \epsilon \\ &\le \sum_{k=1}^{K_j}\sum_{h=1}^{H}\left|\max_{a\in\mathcal{A}} Q_h^*(x_h^k, a) - Q_h^*(x_h^k, y_h^k)\right| + \epsilon \end{aligned} \tag{20}$$

with probability at least $1 - 2\exp\left(-\frac{2\epsilon^2}{K_j H^3}\right)$. We take $\epsilon = 10\sqrt{H^3 K_j \log T}$; then the probability is at least $1 - 2\exp\left(-\frac{200 H^3 K_j \log T}{K_j H^3}\right) = 1 - 2e^{(-200\log T)} = 1 - \frac{2}{T^{200}}$.

On the other hand, let a^* denote the action that achieves $\max_{a\in\mathcal{A}} Q_h^*(x_h^k,a)$, then the first term inside the sum in the right hand side of the Inequality (18) is

$$\begin{aligned}Q_h^*(x_h^k,a^*)-Q_h^*(x_h^k,y_h^k) \le& Q_h^*(x_h^k,a^*)-Q_h^k(x_h^k,a^*)+Q_h^k(x_h^k,a^*)-Q_h^k(x_h^k,y_h^k)+Q_h^k(x_h^k,y_h^k)-Q_h^*(x_h^k,y_h^k)\\ \le&\left|Q_h^*(x_h^k,a^*)-Q_h^k(x_h^k,a^*)\right|+\left(Q_h^k(x_h^k,a^*)-Q_h^k(x_h^k,y_h^k)\right)+\left|Q_h^k(x_h^k,y_h^k)-Q_h^*(x_h^k,y_h^k)\right|\\ \le&\left|Q_h^*(x_h^k,a^*)-Q_h^k(x_h^k,a^*)\right|+\mathrm{CB}_1+\left|Q_h^k(x_h^k,y_h^k)-Q_h^*(x_h^k,y_h^k)\right|\end{aligned} \tag{21}$$

where the last inequality is due to the fact that the second term on the right-hand side is upper-bounded by the confidence interval CB_1 by definition of the running set in Algorithm 1. Recall that $\mathrm{CB}_1 \le \mathcal{O}\Big(\sqrt{H^5\iota}/\sqrt{K_j-1}\Big)$.

On the other hand, by definition of a^*, the left-hand side is non-negative. Therefore, the right-hand side of Equation (21) is also an upper bound on the absolute value of the left-hand side. Therefore, we get that the first term on the right-hand side of Inequality (20) is upper-bounded by:

$$\sum_{k=1}^{K_j}\sum_{h=1}^{H}\left|Q_h^*(x_h^k,a^*)-Q_h^*(x_h^k,y_h^k)\right| \le \sum_{k=1}^{K_j}\sum_{h=1}^{H}\left|Q_h^*(x_h^k,a^*)-Q_h^k(x_h^k,a^*)\right|+\sum_{k=1}^{K_j}\sum_{h=1}^{H}\left|Q_h^k(x_h^k,y_h^k)-Q_h^*(x_h^k,y_h^k)\right|+\mathrm{CB}_1$$

where the first term and the third term are both upper-bounded by $\sum_{k=1}^{K_j}\sum_{h=1}^{H}\max_a\left|Q^*(x_h^k,a)-Q_h^k(x_h^k,a)\right|$.

Let w^* denote the best arm, that is, the arm that correctly chooses (one of) the time steps with the highest optimal base-stock level as the beginning of the cycles. By definition, the best arm has the highest optimal value function for the beginning of its cycles $V_1^{*(w^*)} \equiv \max_w V_1^{*(w)}$, which corresponds to having the highest expected per-episode reward among the arms.

When $w=w^*$, by part 1 of Lemma 1, we know that $\sum_{k=1}^{K_j}\sum_{h=1}^{H}\max_a\left|Q^*(x_h^k,a)-Q_h^k(x_h^k,a)\right|$ is bounded by $\frac{HK_j\delta_h}{\sqrt{K_j-1}}$ with probability at least $1-\frac{1}{A^5T^5}$. Therefore, let $E_{K_j}^{w^*}$ be the (random) total reward for arm w^* after pulling it for K_j cycles, then using the fact that $\delta_h \le 4\sqrt{H^5\iota}$ again, the difference between the expected optimal reward for any arm w and our estimated reward after K_j samples of the arm w is

$$\left|K_jV_1^{*(w^*)}-E_{K_j}^{w^*}\right| \le \mathcal{O}\big(\sqrt{H^7K_j\log T}\big) \tag{22}$$

with probability at least $1-\frac{1}{T^5}$.

Let w_2 denote any suboptimal arm that has not been eliminated before being pulled K_j times. When $w=w_2$, then because of trimming, our estimated reward after K_j could be further lowered:

$$K_jV_1^{*(w_2)} \ge E_{K_j}^{w_2}-\mathcal{O}\big(\sqrt{H^7K_j\log T}\big) \tag{23}$$

with probability at least $1-\frac{1}{T^5}$.

By definition, its optimal value is $V_1^{*(w_2)} \le V_1^{*(w^*)}$. Let $E_{K_j}^{w_2}$ be the total reward for arm w_2 after pulling it for K_j cycles. Then by Equations (22) and (23), after K_j samples, with probability $1-\frac{1}{T^5}$,

$$\begin{aligned}&E_{K_j}^{w^*}+C_2\sqrt{H^7K_j\iota} \ge K_jV_1^{*(w^*)} \ge K_jV_1^{*(w_2)} \ge E_{K_j}^{w_2}-C_2\sqrt{H^7K_j\iota}\\ \implies\ &E_{K_j}^{w^*} \ge E_{K_j}^{w_2}-2C_2\sqrt{H^7K_j\iota}\end{aligned} \tag{24}$$

for the same C_2 we used in the confidence bound CB in Algorithm 2.

Since this holds for all suboptimal arms w_2 and all no more than $\log T$ different values of K_j, by union bound, the probability of *Meta-HQL* never eliminating the best arm is at least $1-\frac{H\log T}{T^5} \le \frac{1}{T^4}$. □

Proof of Theorem 3 Suppose arm w_2 was eliminated after $K_j = K_{j(w_2)}$ samples of arm w_2. Then, arm w_2 was not eliminated when it was pulled $\frac{K_j}{2}$ times. To analyze the regret accumulated from pulling arms, observe that in fact each arm has total regret $\mathcal{O}\Big(\sqrt{\frac{H^7K_j\iota}{2}}\Big)$ with probability at least $1-T^{-5}$ not only on its first $\frac{K_j}{2}$ samples, but also on its $\frac{K_j}{2}+1$ through K_j-th sample, as detailed below.

From the proof of Lemma 4, we know with probability at least $1-\frac{1}{T^4}$,

$$\begin{aligned} E^{w_2}_{\frac{K_j}{2}} \geq E^{w^*}_{\frac{K_j}{2}} - 2C_2\sqrt{\frac{H^7K_j\iota}{2}} \geq \frac{K_jV_1^{*(w^*)}}{2} - \mathcal{O}\Big(\sqrt{\frac{H^7K_j\iota}{2}}\Big) \\ \implies \frac{K_jV_1^{*(w_2)}}{2} \geq \frac{K_jV_1^{*(w^*)}}{2} - \mathcal{O}\Big(\sqrt{H^7K_j\iota}\Big) \end{aligned} \tag{25}$$

Since $E^{w_2}_{K_j} \geq K_jV_1^{*(w_2)} - \mathcal{O}\Big(\sqrt{H^7K_j\iota}\Big)$ by Inequality (6) of Lemma 1, we know that

$$E^{w_2}_{K_j} \geq K_jV_1^{*(w^*)} - \mathcal{O}\Big(\sqrt{H^7K_j\iota}\Big) \tag{26}$$

with probability at least $1-\frac{1}{T^4}$. Therefore, the total regret from playing arm w_2 is

$$K_jV_1^{*(w^*)} - E^{w_2}_{K_j} \leq \mathcal{O}(\sqrt{H^7K_j\iota}) \tag{27}$$

Summing over all suboptimal arms to find the total regret incurred when pulling arms,

$$Regret_{arms} \leq \sum_{w'\in[H]} \mathbf{E}\Big[K_{j(w')}V_1^{*(w^*)} - E^{w'}_{K_{j(w')}}\Big]$$

Since $\sum_{w'\in[H]} K_{j(w')} \leq K$, Jensen's Inequality implies $\sum_{w'\in[H]}\sqrt{K_{j(w')}} \leq H\cdot\sqrt{\frac{K}{H}} = \sqrt{KH}$. Therefore, the total regret incurred by pulling arms is upper-bounded by $\mathcal{O}(\sqrt{H^7K_j\iota})$.

The Regret_{gap} term in the total regret caused by discretization contributes $\mathcal{O}(1/T)$ to the regret. By Lemma 3, switching between arms contributes $\text{Regret}_{switching} = \mathcal{O}(H\gamma\log^2 T)$ to the total regret. The low probability T^{-4} of failure in applying Azuma-Hoeffding has negligible regret contribution. Therefore the main regret term is given by the regret accumulated while pulling arms Regret_{arms}

$$\begin{aligned} \text{Regret}_{total} =& \text{Regret}_{gap} + \text{Regret}_{arms} + \text{Regret}_{switching} \\ \leq& \mathcal{O}(1/T) + \left(\mathcal{O}(\sqrt{H^7T\iota})\times 1 + \mathcal{O}(T^2)\times\frac{1}{T^4}\right) + \mathcal{O}(H\gamma\log^2 T) \\ =& \tilde{\mathcal{O}}\big(\sqrt{H^7T}\big). \quad \square \end{aligned} \tag{28}$$

Appendix D: Assumption of 0 Purchasing Costs

We want to show that for our episodic lost-sales model, we can amortize the unit purchasing costs c_h into unit holding costs o_h and unit lost-sales penalty p_h. First we know that for any $h \geq 2$

$$\begin{aligned} y_h - x_h &= y_h - D_h + D_h - x_h = (y_h - D_h)^+ - (D_h - y_h)^+ + D_h - x_h \\ &= (y_h - D_h)^+ - (D_h - y_h)^+ + D_h - (y_{t-1} - D_{t-1})^+ \end{aligned} \tag{29}$$

Then the total sum of costs starting from time step 2 is

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

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

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

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

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 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.

Similar amortizing works for the single-product backlogged model with zero lead time.

Appendix E: Preliminaries for the episodic multi-product backlogging model

We describe the MDP$(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$ for the multi-product backlogging model in this section. The current state $\mathbf{x}_h \in \mathbb{R}^{n\times L}$ is the concatenation of the current on-hand inventory $\mathbf{I}_h$ and the list of inventories ordered in the pipeline that are still in transit $\mathbf{y}_{h-L+1}, \mathbf{y}_{h-L+2}, \ldots, \mathbf{y}_{h-1}$.

For the multi-product backlogging model, we do a similar transformation[2] on the costs so that the per-period reward of any policy over an episode is bounded by $[0,1]$.

We discretize both the state and action spaces to consist of multiples of $\varepsilon = \frac{M-m}{T^2}$. Rounding all the demands and orders to an adjacent multiple of ε (using a fixed but arbitrary rule) transforms any continuous policy to a discrete policy with at most $\mathcal{O}(H\varepsilon)$ additive error per time-step (due to accumulation over the episode) and hence $\mathcal{O}(H\varepsilon T \times n) = \mathcal{O}(\frac{n(M-m)}{K}) = o(1)$ total additive error in the cost. Note that technically, we might round a tiny order to $\mathbf{0}$, where the reward function is not Lipschitz. However, this only helps as the reward is upper semi-continuous. Therefore solving the discretized problem with regret Regret_{MDP} solves the continuous problem with regret $\text{Regret}_{MDP} + \text{Regret}_{gap} = \text{Regret}_{MDP} + \mathcal{O}(\frac{n(M-m)}{K}) = \text{Regret}_{MDP}$, since $K = \Theta(T)$.

Since the action set for the multi-product backlogging model includes any feasible replenishment amount within the order limits, the reward and leftover inventory depend on both the state and the action. Therefore, we do not simplify the notation $Q(x,y)$ to $Q(x)$.

Appendix F: Regret analysis for *FQL*

For *FQL*, we are able to adopt similar notations and analysis in Jin et al. (2018) (but adapted to our full-feedback setting).

[2] We scale the negated costs down by a factor of $\Theta\big(n \cdot \max(F_h, M|o_h|, M|b_h|)\big)$ and then shift to the right.

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 $Q_h^*(x,y) = \left(r_h + \mathbb{P}_h V_{h+1}^*\right)(x,y)$.

FQL updates the Q values in the following way for any $(x,y)\in\mathcal{A}$ at any time step:

$$Q_h^{k+1}(x,y) \leftarrow (1-\alpha_k) Q_h^k(x,y) + \alpha_k [r_h^{k+1}(x,y) + V_{h+1}^k(x_{h+1})] \tag{30}$$

Then by the definition of weights α_t^k, we have

$$Q_h^k(x,y) = \alpha_{k-1}^0 H + \sum_{j=1}^{k-1} \alpha_{k-1}^j \left[r_h^j(x,y) + V_{h+1}^j\left(x_{h+1}^j\right)\right] \tag{31}$$

The following two lemmas are variations of Lemma 7 and Lemma 8.

LEMMA 11. *For any $(x,y,h,k)\in\mathcal{S}\times\mathcal{A}\times[H]\times[K]$, we have*

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

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 $\left[\hat{\mathbb{P}}_h^i V_{h+1}\right](x,y) := V_{h+1}\left(x_{h+1}^i\right)$, and the fact that $\sum_{i=0}^{k-1} \alpha_{k-1}^i = 1$, we have

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

Subtracting Equation 31 from this equation gives us Lemma 11. □

LEMMA 12. *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:*

$$0 \le \left(Q_h^k - Q_h^*\right)(x,y) \le \alpha_{k-1}^0 H + \sum_{i=1}^{k-1} \alpha_{k-1}^i \left(V_{h+1}^i - V_{h+1}^*\right)\left(x_{h+1}^i\right) + c\sqrt{\frac{H^3\iota}{k-1}} \tag{32}$$

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 ith time. Let $\mathcal{F}_h^i$ be the σ-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 with respect to the filtration $\{\mathcal{F}_h^i\}_{i\ge0}$. Then by Azuma-Hoeffding Theorem, we have that with probability at least $1-p/SAT$:

$$\left|\sum_{i=1}^{k-1} \alpha_k^i \cdot \left[\left(\hat{\mathbb{P}}_h^i - \mathbb{P}_h\right) V_{h+1}^*\right](x,y) + r_h^i - \mathbb{E}[r_h^i]\right| \le \frac{cH}{2}\sqrt{\sum_{i=1}^{k-1}\left(\alpha_{k-1}^i\right)^2 \cdot \iota} \le c\sqrt{\frac{H^3\iota}{k-1}} \tag{33}$$

for some constant c.

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

$$\left|\sum_{i=1}^{k-1} \alpha_{k=1}^i \left[\left(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h\right) V_{h+1}^*\right](x,y) + r_h^i - \mathbb{E}[r_h^i]\right| \le c\sqrt{\frac{H^3\iota}{k-1}} \tag{34}$$

Then the right-hand side of Lemma 12 follows from Lemma 11 and Inequality (34). The left-hand side also follows from Lemma 11 and Inequality (34) using induction on $h = H, H-1, \ldots, 1$. □

Proof of Theorem 4: Define $\Delta_h^k := \left(V_h^k - V_h^{\pi_k}\right)(x_h^k)$ and $\phi_h^k := \left(V_h^k - V_h^*\right)(x_h^k)$.

By Lemma 33, 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:

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

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. Here y_h^k denotes the base-stock levels taken at stage h of episode k, which means $y_h^k = \arg\max Q_h^k(y')$.

$$\begin{aligned}
\Delta_h^k &= \left(V_h^k - V_h^{\pi_k}\right)(x_h^k) \overset{(1)}{\leq} \left(Q_h^k - Q_h^{\pi_k}\right)(x_h^k, y_h^k) \\
&= \left(Q_h^k - Q_h^*\right)(x_h^k, y_h^k) + \left(Q_h^* - Q_h^{\pi_k}\right)(x_h^k, y_h^k) \\
&\overset{(2)}{\leq} \alpha_{k-1}^0 H + \sum_{i=1}^{k-1}\alpha_{k-1}^i \phi_{h+1}^i + c\sqrt{\frac{H^3\iota}{k-1}} + \left[\mathbb{P}_h\left(V_{h+1}^* - V_{h+1}^{\pi_k}\right)\right](x_h^k, y_h^k) \\
&= \alpha_{k-1}^0 H + \sum_{i=1}^{k-1}\alpha_{k-1}^i \phi_{h+1}^i + c\sqrt{\frac{H^3\iota}{k-1}} + \left[\left(\mathbb{P}_h - \hat{\mathbb{P}}_h^k\right)\left(V_{h+1}^* - V_{h+1}^{\pi_k}\right)\right](x_h^k, y_h^k) + (V_{h+1}^* - V_{h+1}^{\pi_k})(x_{h+1}^k) \\
&\overset{(3)}{=} \alpha_{k-1}^0 H + \sum_{i=1}^{k-1}\alpha_{k-1}^i \phi_{h+1}^i + c\sqrt{\frac{H^3\iota}{k-1}} - \phi_{h+1}^k + \Delta_{h+1}^k + \xi_{h+1}^k
\end{aligned} \tag{35}$$

where $\xi_{h+1}^k := \left[\left(\mathbb{P}_h - \hat{\mathbb{P}}_h^k\right)\left(V_{h+1}^* - V_{h+1}^{\pi_k}\right)\right](x_h^k, y_h^k)$ is a martingale difference sequence. Inequality (1) holds because $V_h^k(x_h^k) \leq \max_{\text{feasible } y' \text{ given } x} Q_h^k(x_h^k, y') = Q_h^k(x_h^k, y_h^k)$, and Inequality (2) holds by Lemma 12 and the Bellman equations. Inequality (3) holds by definition $\Delta_{h+1}^k - \phi_{h+1}^k = \left(V_{h+1}^* - V_{h+1}^{\pi_k}\right)\left(x_{h+1}^k\right)$.

In order to compute $\sum_{k=1}^{K}\Delta_1^k$, we need to first bound the first term in Equation 35. 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 35 over the episodes by regrouping:

$$\sum_{k=2}^{K}\sum_{i=1}^{k-1}\alpha_{k-1}^i\phi_{h+1}^i \leq \sum_{i=1}^{K-1}\phi_{h+1}^i\sum_{k=i+1}^{\infty}\alpha_{k-1}^i \leq \sum_{i=1}^{K-1}\phi_{h+1}^i\sum_{k'=i}^{\infty}\alpha_{k'}^i \leq \left(1+\frac{1}{H}\right)\sum_{k=1}^{K}\phi_{h+1}^k \tag{36}$$

where the last inequality uses $\sum_{t=i}^{\infty}\alpha_t^i = 1 + \frac{1}{H}$ for every $i \geq 1$ from Lemma 5.

Plugging the above Equation (36) and $\sum_{k=1}^{K}\alpha_k^0 H \leq H$ back into Equation (35), we have:

$$\begin{aligned}
\sum_{k=1}^{K}\Delta_h^k \leq& H + \sum_{k=2}^{K}\Delta_h^k \\
\leq& H + H + \left(1+\frac{1}{H}\right)\sum_{k=1}^{K}\phi_{h+1}^k - \sum_{k=2}^{K}\phi_{h+1}^k + \sum_{k=2}^{K}\Delta_{h+1}^k + \sum_{k=2}^{K}c\sqrt{\frac{H^3\iota}{k-1}} + \sum_{k=2}^{K}\xi_{h+1}^k \\
\leq& 2H + \phi_{h+1}^1 + \frac{1}{H}\sum_{k=2}^{K}\phi_{h+1}^k + \sum_{k=2}^{K}\Delta_{h+1}^k + \sum_{k=2}^{K}c\sqrt{\frac{H^3\iota}{k-1}} + \sum_{k=2}^{K}\xi_{h+1}^k \\
\leq& 3H + \left(1+\frac{1}{H}\right)\sum_{k=2}^{K}\Delta_{h+1}^k + \sum_{k=2}^{K}c\sqrt{\frac{H^3\iota}{k-1}} + \sum_{k=}^{K}\xi_{h+1}^k
\end{aligned} \tag{37}$$

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

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

where

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

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

$$\left|\sum_{h=1}^{H}\sum_{k=1}^{K}\xi_{h+1}^{k}\right| = \left|\sum_{h=1}^{H}\sum_{k=1}^{K}\left[\left(\mathbb{P}_h - \hat{\mathbb{P}}_h^k\right)\left(V_{h+1}^{*} - V_{h+1}^{\pi_k}\right)\right]\left(x_h^k, y_h^k\right)\right| \le cH\sqrt{T_l} \ \le \tilde{\mathcal{O}}(\sqrt{H^4T}) \tag{38}$$

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

$$\text{Regret}_{total}(K) = \text{Regret}_{MDP}(K) + \text{Regret}_{MDP}(gap) = \mathcal{O}(H^2\sqrt{n(L+1)T\log T}) \tag{39}$$

We multiply the constant $O\big(n\cdot\max(F_h, M|o_h|, M|b_h|)\big)$ back because we previously scaled the costs to have the reward for each time period bounded by 1. This yields a $\mathcal{O}(H^2n\sqrt{n(L+1)}\cdot\max(F_h, M|o_h|, M|b_h|))$ total dependence on setting parameters for our $\tilde{\mathcal{O}}(T)$ regret. □

When $L=0, n=1, F_h=0$, the total regret of *FQL* on the single-product backlogging model with a lead time and an order limit is $\tilde{\mathcal{O}}(\sqrt{T})$ with an $\mathcal{O}\big(H^2M\max(|o_h|,|b_h|)\big)$ dependence on all constant parameters. This is smaller than the dependence of *HQL* applied on the single-product backlogging model with a lead time and an order limit by a factor of H.

Appendix G: Regret analysis for *MimicQL*

Proof of Theorem 5: Let ℓ denote the maximum order limit. Let o denote the maximum unit holding cost. The expected amount of time until synchronization is no more than $\mathcal{O}(nL\ell\gamma)$. During each of these time steps the holding cost is a constant nMo. Then the additional cost *Mimic-FQL* incurs each time by not discarding is bounded by $\mathcal{O}(n^2\ell ML\gamma o)$. This is a constant term, but since discarding happens at the end of every episode, the total additional regret incurred is $\mathcal{O}(Kn^2\ell ML\gamma o)$. Note that this term is linear in T when $K=\Theta(T)$, and we will perform additional techniques to obtain a total regret that is sublinear in T.

With positive lead time, the optimal policy OPT in the non-discarding model can have a larger or smaller total expected cost than the optimal policy $\underline{OPT}$ for the intermediate MDP. We want to show that Cost of the optimal policy OPT in the non-discarding model, will not be too much lower than the $\underline{\text{Cost}}$ of the optimal policy $\underline{OPT}$ for the intermediate MDP.

Consider a policy π_1 on the intermediate MDP. At the beginning of each episode, π_1 starts with zero inventory and zero replenishment because discarding at the end of the previous episode. In the first L time steps of the second episode, π_1 orders the replenishment in a way that it ends up with the same inventory vector and replenishment vector as OPT at the end of L time steps. Starting at time step $(L+1)$, π_1 completely copies OPT under the beginning of the next episode, where π_1 starts with zero inventory and replenishment again. For each episode, the cost of this policy π_1 is at most $\mathcal{O}\Big(LF+LnMc\Big)$ more than the cost of OPT. Therefore, the total expected cost of π_1 is at most $\mathcal{O}\Big(KLF+KLnMc\Big)$ more than the total expected cost of OPT. On the other hand, since by definition, the total expected cost of the optimal policy on the intermediate MDP is no higher than the total expected cost of π_1, we know that the total expected cost of $\underline{OPT}$ is at most $\mathcal{O}\Big(KLF+KLnMc\Big)$ more than the total expected cost of OPT.

So far we have argued that in the case of zero lead time,

$$\begin{aligned}&1.\ \underline{Cost}_T(\underline{OPT}) - \mathcal{O}\Big(KLF + KLnMc\Big) \le \text{Cost of } OPT\\ &2.\ \underline{\text{Cost}} \text{ of } FQL \le \text{Cost of } \textit{Mimic-FQL} \le \underline{\text{Cost}} \text{ of } FQL + \mathcal{O}\big(Kn^2\ell ML\gamma o\big)\end{aligned} \tag{40}$$

Then we know that

$$\begin{aligned}\text{Regret}_{Mimic-FQL} :=& \text{Cost of } \textit{Mimic-FQL} - \text{Cost of } OPT\\ \le& \text{Cost of } \textit{Mimic-FQL} - \text{Cost of } \underline{OPT} + \text{Cost of } \underline{OPT} - \text{Cost of } OPT\\ \le& \text{Cost of } \textit{Mimic-FQL} - \text{Cost of } \underline{OPT} + \mathcal{O}\Big(KLF + KLnMc\Big)\\ \le& \Big(\text{Cost of } \textit{Mimic-FQL} - \underline{\text{Cost}} \text{ of } FQL\Big) + \Big(\underline{\text{Cost}} \text{ of } FQL - \underline{\text{Cost}} \text{ of } \underline{OPT}\Big)\\ &+ \mathcal{O}\Big(KLF + KLnMc\Big)\\ \le& \mathcal{O}\big(Kn^2\ell ML\gamma o\big) + \tilde{\mathcal{O}}\big(J^2\sqrt{T}\big) + \mathcal{O}\Big(KLF + KLnMc\Big)\end{aligned} \tag{41}$$

where we recall that the second term on the last line is bounded by Theorem 1 for the episodic model.

Since $K := \frac{T}{J}$, we know that $\text{Regret}_{Mimic-FQL} \le \mathcal{O}\big(\frac{T(n^2\ell ML\gamma o + LF + LnMc)}{J}\big) + \tilde{\mathcal{O}}\big(J^2\sqrt{T}\big)$. Choosing J to be a multiple of H of size $J = \Theta(T^{1/6})$ now yields the regret bound $\tilde{\mathcal{O}}\big(T^{5/6}\big)$. Note that $T^{1/6}$ might not be an integer multiple of H, then we take J to be the closest multiple of H to $T^{1/6}$. □

Appendix H: General definitions and assumptions for wider application of our policies

H.1. Full Feedback.

The formal definition of *full feedback* is as follows. Immediately after taking an action a_t at time t, once the environmental randomness D_t is realized, the agent learns what the counterfactual reward $r_t(s,a)$ and next state $s_{t+1}(s,a)$ would have been for all feasible state action pairs $(s,a) \in \mathcal{S} \times \mathcal{A}$ for that specific time step t.

We notice this feedback structure in the backlogging inventory problems: in the backlogging model, we observe the actual realized demand, which allows us to deduce what the cost and leftover inventory would be for any action. Trivially, the backlogging model also possesses one-sided feedback.

For problems that possess the full-feedback structure, *FQL* is applicable with our regret bound guarantee.

H.2. One-Sided Feedback.

The formal definition of *one-sided feedback* is as follows. Immediately after taking an action a at step t, once the environmental randomness D_t is realized, we learn what the reward and next state would have been if any action that lie on *one side* of a is taken, i.e., all $a' \le a$ for the *lower*-sided-feedback structure for that specific time step t (or all $a' \ge a$ for the *higher*-sided-feedback structure). This implies that the action space can be embedded in a compact subset of $\mathbb{R}$.

We notice this feedback structure is in the lost-sales inventory control problem: once the demand D_t is realized for that time step, if the demand is lower than our chosen base-stock level y_t, we will observe the actual D_t; otherwise we will observe the $\min(y_t, D_t)$ part of the demand, which lets us deduce what the pseudo-cost and leftover inventory would be if the agent had taken any action (base-stock level) lower than y_t. Mathematically, for any $y'_t \le y_t$, $\min(y'_t, D_t) = \min(y'_t, \min(y_t, D_t))$.

We list a number of assumptions that need to hold for the *lower*-sided-feedback setting. In the case of the *higher*-sided-feedback setting, Assumptions 2 and 3 would be symmetric to Assumptions 2 and 3 below. If the set of feasible actions at any time is unaffected by the current state, then the assumptions below are unnecessary. However, in that case, even though our algorithm still applies, the MDP problem can be reduced to a number of bandit problems.

Assumptions (lower-sided):

1. The optimal Q-value functions are concave.
2. The current feasible action set at time t is of the form $\mathcal{A} \cap [a, \infty)$, for some $a \in \mathbb{R}$ non-decreasing in x_t.
3. Conditioned on the environmental randomness, the next state $x_{t+1}(\cdot)$ is non-decreasing in y_t.
4. The reward and transition only depend on the action, the time step and the environmental randomness, even though the feasible action set can depend on the state. So $Q(x, y)$ can be simplified to $Q(y), \forall y$ feasible for x.

These assumptions impose a specific structure on the problem, which is often satisfied in important OR and finance problems, e.g. inventory control, portfolio management, airline's overbook policy, online second price auctions, etc. See an overview of these applications in Section 8.