Electronic Companion

"On the Hardness of Learning from Censored and Nonstationary Demand"

Proofs of Main Results

Let $k \in \mathbb{N}$, with $k \leq D$, and denote by $\mathbb{L}_k$ the $k \times k$ identity matrix, and by $\mathbb{M}_k$ the $k \times (D-k)$ matrix, where $\mathbb{M}_k(i,j) = 1$ if $i = k$, and 0 otherwise. Finally, let e_d be the $(D+1)$-dimensional column vector, with $e_d(j) = 1$ if $j = d$, and 0 otherwise.

The *signal matrix* of inventory decision $i \in \mathcal{I}$, denoted by $\mathbb{S}_i$, is a $(i+1) \times (D+1)$ matrix, where element $\mathbb{S}_i(k,j) = 1$ if the sales of the firm are equal to $k \in \{0,1,\ldots,i\}$, assuming that the demand is equal to $j \in \mathcal{D}$ and the inventory is equal to i, and 0 otherwise. It can be verified that $\mathbb{S}_i$ is equal to the concatenated matrix $\mathbb{S}_i = \left[\mathbb{L}_{i+1} \,|\, \mathbb{M}_{i+1}\right]$.

Lemma 1: Local Observability

Lemma 1. *Let $i, j \in \mathcal{I}$ be arbitrary inventory decisions. There exist vectors $v_i \in \mathcal{R}^{i+1}$ and $v_j \in \mathcal{R}^{j+1}$ such that*

$$\left(v_i^T \mathbb{S}_i - v_j^T \mathbb{S}_j\right) e_d = c(i,d) - c(j,d), \qquad d \in \mathcal{D}.$$

Proof. Consider the $(i+1)$-dimensional column vector v_i, where

$$v_i(k) = hi - (h+b)(k-1), \qquad k \in \{1,2,\ldots,i+1\}.$$

Define similarly the vector v_j. The result follows through straightforward calculations. ■

Hence, the game between the inventory manager and the market is locally observable, in the sense of Definition 6 in Bartók et al. [2014]. This classifies the repeated newsvendor problem with demand learning via censored data as an "easy" partial monitoring problem (see Section 2.1), which implies that the correct scaling of the expected regret is $\Theta\left(\sqrt{T}\right)$.

Lemma 2: Bias and Variance of the Estimator

Lemma 2. *The cost estimator in Eq.* (5) *satisfies:*

$$\mathbb{E}_t\left[\widetilde{c}(i,d_t)\right] = v_i^T \mathbb{S}_i e_{d_t} + \beta,$$

so that $\mathbb{E}_t\left[\widetilde{c}(i,d_t)\right] \in (0,2\beta)$*, and*

$$\mathbb{E}_t\left[\widetilde{c}(i,d_t)^2\right] \leq \frac{4\beta^2}{\mathbb{P}_t(I_t \geq i)},$$

where $\mathbb{P}_t(\cdot)$ *and* $\mathbb{E}_t[\cdot]$ *respectively denote the probability and the expectation conditioned on the history of interaction up until the beginning of round* t*.*

Proof. The first part of the lemma follows directly by noting that

$$\mathbb{E}_t\left[\widetilde{c}(i,d_t)\right] = \frac{\mathbb{E}_t\left[\mathbb{I}_{\{I_t \geq i\}}\right]}{\mathbb{P}_t(I_t \geq i)}\left(v_i^T \mathbb{S}_i e_{d_t} + \beta\right) = v_i^T \mathbb{S}_i e_{d_t} + \beta.$$

The fact that $\mathbb{E}_t\left[\widetilde{c}(i,d_t)\right] \in (0,2\beta)$ is a direct consequence of β being an upper bound on the absolute value of $v_i^T S_i e_{d_t}$, for every $i \in \mathcal{I}$ and d_t. Hence, regarding the second part of the lemma, we have that

$$\mathbb{E}_t\left[\widetilde{c}(i,d_t)^2\right] = \frac{\mathbb{E}_t\left[\mathbb{I}_{\{I_t \geq i\}}\right]}{\mathbb{P}_t(I_t \geq i)^2}\left(v_i^T \mathbb{S}_i e_{d_t} + \beta\right)^2 \leq \frac{4\beta^2}{\mathbb{P}_t(I_t \geq i)}.$$

■

Note that the proposed estimator is biased, as $\mathbb{E}_t\left[\widetilde{c}(i,d_t)\right] \neq c(i,d_t)$. In particular, the estimator is pessimistic in the sense that it always overestimates the actual cost incurred. A direct corollary of Lemmas 1 and 2 is the following result.

Lemma 3: Unbiased Estimator

Lemma 3. *The cost estimator in Eq.* (5) *is unbiased when inferring the difference in cost between two actions:*

$$\mathbb{E}_t\left[\widetilde{c}(i,d_t) - \widetilde{c}(j,d_t)\right] = c(i,d_t) - c(j,d_t), \qquad i,j \in \mathcal{I}.$$

The significance of Lemma 3 lies in the fact that the regret, by definition, is a metric that is based on cost differences. This facilitates the analysis in our main result, which characterizes the performance of the proposed inventory management policy with respect to the regret criterion.

Lemma 4

Lemma 4. *Let* $\mathbf{p} = (p_1, p_2, \ldots, p_N)$ *be any probability distribution satisfying* $p_i \geq \omega > 0$ *for all* $i \in \{1, 2, \ldots, N\}$. *Then, the following inequality is satisfied:*

$$\sum_{i=1}^{N} \frac{p_i}{\sum_{j=i}^{N} p_j} \leq 5 \log(1 + 1/\omega) + 3.$$

Proof. We fix $\omega > 0$ and partition the actions into the sets $M > 1$ sets, with the m-th set holding probability mass of at least 2^m for all m. Precisely, we let $k_0 = N + 1$ and define k_m for $m > 0$ recursively as

$$k_{m+1} = \max \left\{ i : \sum_{j=i}^{k_m - 1} p_j \geq 2^m \omega \right\},$$

with $\max\{\emptyset\} = 1$, and we let $M = \min\{m : k_m = 1\}$. Then, we set $\mathcal{I}_m = [k_m, k_{m-1}]$ for all $m = 1, \ldots, M$. Notice that, for any $m < M$, this choice guarantees that

$$\sum_{j=k_m}^{N} p_j = \sum_{n=0}^{m-1} \sum_{j=k_{n+1}}^{k_n - 1} p_j \geq \sum_{n=0}^{m-1} 2^n \omega = \left(2^m - 1\right) \omega,$$

which can be seen to imply an upper bound on M since

$$1 = \sum_{j=k_M}^{N} p_j \geq \sum_{j=k_{M-1}}^{N} p_j \geq \left(2^{M-1} - 1\right) \omega \geq 2^{M-1} \omega.$$

Indeed, this implies that $M \leq \log_2(1 + 1/\omega) + 1$. Furthermore, we can write the following:

$$\begin{aligned}
\sum_{i=1}^{N} \frac{p_i}{\sum_{j=i}^{N} p_j} &= \sum_{m=0}^{M-1} \sum_{i=k_{m+1}}^{k_m - 1} \frac{p_i}{\sum_{j=i}^{N} p_j} \\
&= \sum_{m=0}^{M-1} \left(\frac{p_{k_{m+1}}}{\sum_{j=k_{m+1}}^{N} p_j} + \sum_{i=k_{m+1}+1}^{k_m - 1} \frac{p_i}{\sum_{j=i}^{N} p_j} \right) \\
&\leq \sum_{m=0}^{M-1} \left(1 + \sum_{i=k_{m+1}+1}^{k_m - 1} \frac{p_i}{\sum_{j=k_m}^{N} p_j} \right) \\
&\leq \sum_{m=0}^{M-1} \left(1 + \sum_{i=k_{m+1}+1}^{k_m - 1} \frac{p_i}{(2^m - 1)\,\omega} \right) \\
&\leq \sum_{m=0}^{M-1} \left(1 + \frac{2^m \omega}{(2^m - 1)\,\omega} \right) = \sum_{m=0}^{M-1} \frac{2^m + 2^m - 1}{2^m - 1} = \sum_{m=0}^{M-1} \frac{2^{m+1} - 1}{2^m - 1} \leq 3M,
\end{aligned}$$

where we used the fact that $\sum_{i=k_{m+1}+1}^{k_m - 1} p_i \leq 2^m \omega$ holds by definition of the index k_m. The proof is concluded by putting this together with the upper bound on M and bounding $3/\log(2) \leq 5$. ■

Proof of Theorem 1

Our analysis largely follows the steps of the proof of Theorem 3.1 by Auer et al. [2002], combined with our Lemmas 2 and 3. We analyze a slightly more general version of the EWF algorithm that uses an arbitrary exploration distribution μ, with μ_i being the probability of taking action i in exploration rounds. More precisely, we consider a version of EWF that computes its probability distributions over the actions as

$$p_i(t) = (1-\gamma)\frac{W_i(t-1)}{W(t-1)} + \gamma\mu_i, \qquad i \in \mathcal{I}.$$

The statement will follow from setting $\mu_i = 1/N$ for all actions i.

The key idea of the analysis is studying the term $\log\big(W(T)/W(0)\big)$ which, as we show shortly, relates closely to the regret. We start by constructing a lower bound:

$$\begin{aligned}
\log\left(\frac{W(T)}{W(0)}\right) &= \log\big(W(T)\big) - \log\big(W(0)\big) \\
&= \log\left(\sum_{i\in\mathcal{I}} W_i(T)\right) - \log\left(\sum_{i\in\mathcal{I}} W_i(0)\right) \\
&= \log\left(\sum_{i\in\mathcal{I}} e^{-\eta\widetilde{C}_i(T)}\right) - \log N \\
&\geq \log\left(e^{-\eta\widetilde{C}_{i^*}(T)}\right) - \log N \\
&= -\eta\sum_{t\in\mathcal{T}} \widetilde{c}(i^*, d_t) - \log N, \qquad (10)
\end{aligned}$$

where i^* is the best fixed action in hindsight, for the particular demand sequence.

Then, we derive an upper bound on $\log\big(W(T)/W(0)\big)$:

$$\begin{aligned}
\log\left(\frac{W(T)}{W(0)}\right) &= \log\left(\prod_{t\in\mathcal{T}} \frac{W(t)}{W(t-1)}\right) \\
&= \sum_{t\in\mathcal{T}} \log\left(\frac{W(t)}{W(t-1)}\right) \\
&= \sum_{t\in\mathcal{T}} \log\left(\sum_{i\in\mathcal{I}} \frac{W_i(t)}{W(t-1)}\right) \\
&= \sum_{t\in\mathcal{T}} \log\left(\sum_{i\in\mathcal{I}} \frac{W_i(t-1)}{W(t-1)} e^{-\eta\widetilde{c}(i,d_t)}\right) \quad \text{(by Eq. (3))}.
\end{aligned}$$

Note that $e^{-x} \leq 1 - x + x^2/2$, for all $x \geq 0$, and our estimators for the cost of the different decisions are

nonnegative. Thus,

$$\log\left(\frac{W(T)}{W(0)}\right) \leq \sum_{t\in\mathcal{T}} \log\left(\sum_{i\in\mathcal{I}} \frac{W_i(t-1)}{W(t-1)}\left(1-\eta\widetilde{c}(i,d_t)+\frac{\eta^2}{2}\widetilde{c}(i,d_t)^2\right)\right)$$
$$= \sum_{t\in\mathcal{T}} \log\left(1-\eta\sum_{i\in\mathcal{I}} \frac{W_i(t-1)}{W(t-1)}\widetilde{c}(i,d_t)+\frac{\eta^2}{2}\sum_{i\in\mathcal{I}} \frac{W_i(t-1)}{W(t-1)}\widetilde{c}(i,d_t)^2\right).$$

Moreover, $\log(1+x) \leq x$, for all $x > -1$. Since this is the case with the right-hand side of the expression above (being an upper bound to a sum of exponential terms), we have, using Eq. (4), that

$$\log\left(\frac{W(T)}{W(0)}\right) \leq \sum_{t\in\mathcal{T}} \left(-\eta\sum_{i\in\mathcal{I}} \frac{W_i(t-1)}{W(t-1)}\widetilde{c}(i,d_t)+\frac{\eta^2}{2}\sum_{i\in\mathcal{I}} \frac{W_i(t-1)}{W(t-1)}\widetilde{c}(i,d_t)^2\right)$$
$$= \sum_{t\in\mathcal{T}} \left(-\eta\sum_{i\in\mathcal{I}} \frac{p_i(t)-\gamma\mu_i}{1-\gamma}\widetilde{c}(i,d_t)+\frac{\eta^2}{2}\sum_{i\in\mathcal{I}} \frac{p_i(t)-\gamma\mu_i}{1-\gamma}\widetilde{c}(i,d_t)^2\right). \tag{11}$$

Eqs. (10) and (11) imply that

$$\frac{\eta}{1-\gamma}\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}} p_i(t)\widetilde{c}(i,d_t) - \eta\sum_{t\in\mathcal{T}} \widetilde{c}(i^*,d_t) \leq \log N + \frac{\eta\gamma}{1-\gamma}\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}} \mu_i\widetilde{c}(i,d_t) + \frac{\eta^2}{2(1-\gamma)}\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}} p_i(t)\widetilde{c}(i,d_t)^2.$$

Since $\widetilde{c}(i^*,d_t) \geq 0$, for all $t \in \mathcal{T}$, we have that

$$\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}} p_i(t)\widetilde{c}(i,d_t) - \sum_{t\in\mathcal{T}} \widetilde{c}(i^*,d_t) \leq \frac{\log N}{\eta} + \gamma\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}} \mu_i\widetilde{c}(i,d_t) + \frac{\eta}{2}\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}} p_i(t)\widetilde{c}(i,d_t)^2.$$

This further implies that

$$\mathbb{E}\left[\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}} p_i(t)\big(\widetilde{c}(i,d_t)-\widetilde{c}(i^*,d_t)\big)\right] \leq \frac{\log N}{\eta} + \gamma\mathbb{E}\left[\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}} \mu_i\widetilde{c}(i,d_t)\right] + \frac{\eta}{2}\mathbb{E}\left[\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}} p_i(t)\widetilde{c}(i,d_t)^2\right].$$

The tower rule of expectations along with Eq. (2) and Lemma 3 imply that

$$\mathbb{E}\left[\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}} p_i(t)\big(\widetilde{c}(i,d_t)-\widetilde{c}(i^*,d_t)\big)\right] = \mathbb{E}\left[\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}} p_i(t)\mathbb{E}_t\big[\widetilde{c}(i,d_t)-\widetilde{c}(i^*,d_t)\big]\right]$$
$$= \mathbb{E}\left[\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}} p_i(t)\big(c(i,d_t)-c(i^*,d_t)\big)\right] = \mathbb{E}[\mathcal{R}(T)].$$

Combined with Lemma 2, and the fact that $\mu_i = 1/N$, we get:

$$\mathbb{E}[\mathcal{R}(T)] \leq \frac{\log N}{\eta} + \frac{\gamma}{N}\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}} 2\beta + \frac{\eta}{2}\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}} p_i(t)\frac{4\beta^2}{\mathbb{P}_t(I_t \geq i)}$$

$$= \frac{\log N}{\eta} + 2\beta\gamma T + 2\beta^2\eta \sum_{t\in\mathcal{T}} \sum_{i\in\mathcal{J}} \frac{p_i(t)}{\mathbb{P}_t(I_t \geq i)}.$$

The final step in the proof is to bound the term $\sum_{i\in\mathcal{J}} p_i(t)/\mathbb{P}_t(I_t \geq i)$. This can be directly bounded by applying Lemma 4 with $\omega = \frac{\gamma}{N}$, which gives

$$\sum_{i\in\mathcal{J}} \frac{p_i(t)}{\mathbb{P}_t(I_t \geq i)} = \sum_{i\in\mathcal{J}} \frac{p_i(t)}{\sum_{j=i}^{N} p_j(t)} \leq 5\left(\log\left(\frac{N}{\gamma}+1\right)+1\right) \leq 5\log\left(\frac{3N}{\gamma}+3\right).$$

Consequently,

$$\mathbb{E}[\mathcal{R}(T)] \leq \frac{\log N}{\eta} + 2\beta\gamma T + 10\beta^2\eta T \log\left(\frac{3N}{\gamma}+3\right).$$

By choosing

$$\eta = \sqrt{\frac{\log N}{10\beta^2 T \log\left(\frac{3N}{\gamma}+3\right)}},$$

and bounding $2\sqrt{10} \leq 7$, we have that

$$\mathbb{E}[\mathcal{R}(T)] \leq 7\beta\sqrt{T\log N \log\left(\frac{3N}{\gamma}+3\right)} + 2\beta\gamma T.$$

Finally, by setting $\gamma = 1/\left(2\beta T\right)$, we get:

$$\mathbb{E}[\mathcal{R}(T)] \leq 7\beta\sqrt{T\log N \log\left(6\beta T N + 3\right)} + 1.$$

This concludes the proof.

Proof of Theorem 2

We follow the steps of the proof of Theorem 8.1 in Auer et al. [2002]. To this end, fix a comparator sequence $i_{[T]} \in \mathcal{I}_S^T$, and partition the interval $[1,T]$ into a number of subintervals $I_1 = [1, T_1]$, $I_2 = [T_1+1, T_2]$, …, $I_C = [T_{C-1}+1, T]$, such that i_t remains constant within each interval. Since $i_{[T]} \in \mathcal{I}_S^T$, we have that $C \leq S$. In the remainder of the proof, we bound the regret within each interval, and then combine the obtained bounds to prove a guarantee about the (global) tracking regret.

Fix an arbitrary interval I_s, $s \in \{1,\ldots,C\}$, and let j_s be the action taken during that interval (i.e., $i_t = j_s$, for all $t \in I_s$). Also, let $\tau_s = T_s - T_{s-1}$ be the length of I_s. As in the proof of Theorem 1, we study the term

$\log\big(W(T_s)/W(T_{s-1})\big)$. First, note that

$$\begin{aligned} W_{j_s}(T_s) &\geq W_{j_s}(T_{s-1}+1)\exp\left(-\eta\sum_{t=T_{s-1}+2}^{T_s}\tilde{c}(j_s,d_t)\right) \\ &\geq \frac{\alpha}{N}W(T_{s-1})\exp\left(-\eta\sum_{t=T_{s-1}+2}^{T_s}\tilde{c}(j_s,d_t)\right) \\ &\geq \frac{\alpha}{N}W(T_{s-1})\exp\left(-\eta\sum_{t=T_{s-1}+1}^{T_s}\tilde{c}(j_s,d_t)\right), \end{aligned}$$

where the first and second inequalities follow from the (recursive) definition of the sequence of weights $W_j(t)$, and the last one is a consequence of the non-negativity of the loss estimates $\tilde{c}(i,d_t)$. Hence,

$$\log\left(\frac{W(T_s)}{W(T_{s-1})}\right) \geq \log\left(\frac{W_{j_s}(T_s)}{W(T_{s-1})}\right) \geq \log\left(\frac{\alpha}{N}\right) - \eta\sum_{t=T_{s-1}+1}^{T_s}\tilde{c}(j_s,d_t).$$

On the other hand, we have that

$$\frac{W(t+1)}{W(t)} = \sum_{i\in\mathcal{I}}\frac{W_i(t)e^{-\eta\tilde{c}(i,d_t)}+\frac{\alpha}{N}W(t)}{W(t)} = \sum_{i\in\mathcal{I}}\frac{W_i(t)e^{-\eta\tilde{c}(i,d_t)}}{W(t)} + \alpha,$$

so by a similar line of reasoning as in the proof of Theorem 1, it can be verified that

$$\log\frac{W(T_s)}{W(T_{s-1})} \leq \sum_{t=T_{s-1}+1}^{T_s}\left(-\eta\sum_{i\in\mathcal{I}}\frac{p_i(t)-\gamma/N}{1-\gamma}\tilde{c}(i,d_t) + \frac{\eta^2}{2}\sum_{i\in\mathcal{I}}\frac{p_i(t)-\gamma/N}{1-\gamma}\tilde{c}(i,d_t)^2 + \alpha\right).$$

Combining the two bounds together, taking expectations, and appealing to Lemma 2, we obtain:

$$\mathbb{E}\left[\sum_{t=T_{s-1}+1}^{T_s}\sum_{i\in\mathcal{I}}p_i(t)\big(\tilde{c}(i,d_t)-\tilde{c}(j_s,d_t)\big)\right] \leq \frac{\log\left(\frac{N}{\alpha}\right)}{\eta} + 2\beta\gamma\tau_s + 2\beta^2\eta\mathbb{E}\left[\sum_{t=T_{s-1}+1}^{T_s}\sum_{i\in\mathcal{I}}\frac{p_i(t)}{\mathbb{P}_t(I_t\geq i)}\right] + \alpha\tau_s. \quad (12)$$

As in the proof of Theorem 1, the third term on the right-hand side can be bounded from above using Lemma 4 as follows:

$$\sum_{i\in\mathcal{I}}\frac{p_i(t)}{\mathbb{P}_t(I_t\geq i)} \leq 5\log\left(\frac{3N}{\gamma}+3\right).$$

Using this bound, we can add over all intervals $s\in\{1,\ldots,C\}$ both sides of Eq. (12), and use Lemma 3 to obtain:

$$\mathbb{E}\left[\sum_{t=1}^{T}\big(c(I_t,d_t)-c(i_t,d_t)\big)\right] \leq \frac{S\log\left(\frac{N}{\alpha}\right)}{\eta} + 2\beta\gamma T + 10\beta^2\eta T\log\left(\frac{3N}{\gamma}+3\right) + \alpha T.$$

The statement of the theorem follows from taking the supremum over all $i_{[T]}\in\mathcal{I}_S^T$, and substituting for

the chosen values of γ, η, and α. We note that supremum and expectation can be interchanged in our case, since the comparator sequence, that the supremum is taken with respect to, is deterministic. This would not have been the case, e.g., if the firm competed against an adaptive adversary.

Lemma 5: Variance of Estimator for the Combinatorial Case

Lemma 5. *The second moment of the estimator defined in Equation* (7) *satisfies*

$$\sum_{(i^{(1)},\dots,i^{(K)})\in\mathcal{A}_r} p_{(i^{(1)},\dots,i^{(K)})}(t)\mathbb{E}_t\left[\widetilde{c}\left(i^{(1)},\dots,i^{(K)},r,d_t^{(1)},\dots,d_t^{(K)}\right)^2\right] \leq 20K^2\beta^2\log\left(\frac{3KN}{\gamma}+3\right)+2(f+K\beta)^2.$$

Proof. For simplicity, let us introduce the notation

$$\ell_k\left(i^{(k)},d_t^{(k)}\right)=v_i^T\mathbb{S}_i e_{d_t^{(k)}}+f^{(k)}\mathbb{I}_{\{i^{(k)}>0\}}+\beta,$$

so that each retailer's cost estimate can be written as

$$\widetilde{c}_k\left(i^{(k)},d_t^{(k)}\right)=\frac{\mathbb{I}_{\left\{I_t^{(k)}\geq i^{(k)}\right\}}\ell_k\left(i^{(k)},d_t^{(k)}\right)}{\mathbb{P}_t\left(I_t^{(k)}\geq i^{(k)}\right)},$$

for all $k\in\mathcal{K}$. Also, let $\tilde{I}_t$ be an independent copy of I_t. With this notation, we have that

$$\begin{aligned}
&\sum_{(i^{(1)},\dots,i^{(K)})\in\mathcal{A}_r} p_{(i^{(1)},\dots,i^{(K)})}(t)\mathbb{E}_t\left[\widetilde{c}\left(i^{(1)},\dots,i^{(K)},r,d_t^{(1)},\dots,d_t^{(K)}\right)^2\right]\\
=&\mathbb{E}_t\left[\left(c_0\left(\tilde{I}_t^{(1)},\dots,\tilde{I}_t^{(K)},r\right)+\sum_{k\in\mathcal{K}}\widetilde{c}_k\left(\tilde{I}_t^{(k)},d_t^{(k)}\right)\right)^2\right]\\
\leq&2\mathbb{E}_t\left[c_0\left(\tilde{I}_t^{(1)},\dots,\tilde{I}_t^{(K)},r\right)^2+\left(\sum_{k\in\mathcal{K}}\widetilde{c}_k\left(\tilde{I}_t^{(k)},d_t^{(k)}\right)\right)^2\right],
\end{aligned}$$

where the last step follows from the inequality $(a+b)^2\leq 2\left(a^2+b^2\right)$, which holds for all $a,b\in\mathbb{R}$. The first term can be trivially bounded by $(f+K\beta)^2$. Regarding the second term, we have that

$$\mathbb{E}_t\left[\left(\sum_{k\in\mathcal{K}}\widetilde{c}_k\left(\tilde{I}_t^{(k)},d_t^{(k)}\right)\right)^2\right]$$

$$
\begin{aligned}
&= \mathbb{E}_t \left[\left(\sum_{j \in \mathcal{K}} \frac{\mathbb{I}_{\left\{ I_t^{(j)} \geq \tilde{I}_t^{(j)} \right\}}}{\mathbb{P}_t \left(I_t^{(j)} \geq \tilde{I}_t^{(j)} \right)} \ell_j \left(\tilde{I}_t^{(j)}, d_t^{(j)} \right) \right) \cdot \left(\sum_{k \in \mathcal{K}} \frac{\mathbb{I}_{\left\{ I_t^{(k)} \geq \tilde{I}_t^{(k)} \right\}}}{\mathbb{P}_t \left(I_t^{(k)} \geq \tilde{I}_t^{(k)} \right)} \ell_k \left(\tilde{I}_t^{(k)}, d_t^{(k)} \right) \right) \right] \\
&= \mathbb{E}_t \left[\sum_{j \in \mathcal{K}} \sum_{k \in \mathcal{K}} \frac{\mathbb{I}_{\left\{ I_t^{(j)} \geq \tilde{I}_t^{(j)} \right\}} \mathbb{I}_{\left\{ I_t^{(k)} \geq \tilde{I}_t^{(k)} \right\}}}{\mathbb{P}_t \left(I_t^{(j)} \geq \tilde{I}_t^{(j)} \right) \mathbb{P}_t \left(I_t^{(k)} \geq \tilde{I}_t^{(k)} \right)} \ell_j \left(\tilde{I}_t^{(j)}, d_t^{(j)} \right) \ell_k \left(\tilde{I}_t^{(k)}, d_t^{(k)} \right) \right] \\
&\leq \frac{1}{2} \mathbb{E}_t \left[\sum_{j \in \mathcal{K}} \sum_{k \in \mathcal{K}} \left(\frac{1}{\mathbb{P}_t \left(I_t^{(j)} \geq \tilde{I}_t^{(j)} \right)^2} + \frac{1}{\mathbb{P}_t \left(I_t^{(k)} \geq \tilde{I}_t^{(k)} \right)^2} \right) \mathbb{I}_{\left\{ I_t^{(j)} \geq \tilde{I}_t^{(j)} \right\}} \mathbb{I}_{\left\{ I_t^{(k)} \geq \tilde{I}_t^{(k)} \right\}} \ell_j \left(\tilde{I}_t^{(j)}, d_t^{(j)} \right) \ell_k \left(\tilde{I}_t^{(k)}, d_t^{(k)} \right) \right],
\end{aligned}
$$

again using $(a+b)^2 \leq 2\left(a^2+b^2\right)$. We further have that

$$
\begin{aligned}
&\mathbb{E}_t \left[\left(\sum_{k \in \mathcal{K}} \widetilde{c}_k \left(\tilde{I}_t^{(k)}, d_t^{(k)} \right) \right)^2 \right] \\
&= \mathbb{E}_t \left[\sum_{j \in \mathcal{K}} \sum_{k \in \mathcal{K}} \frac{1}{\mathbb{P}_t \left(I_t^{(j)} \geq \tilde{I}_t^{(j)} \right)^2} \mathbb{I}_{\left\{ I_t^{(j)} \geq \tilde{I}_t^{(j)} \right\}} \mathbb{I}_{\left\{ I_t^{(k)} \geq \tilde{I}_t^{(k)} \right\}} \ell_j \left(\tilde{I}_t^{(j)}, d_t^{(j)} \right) \ell_k \left(\tilde{I}_t^{(k)}, d_t^{(k)} \right) \right],
\end{aligned}
$$

due to the symmetry between j and k, which implies that

$$
\begin{aligned}
&\mathbb{E}_t \left[\left(\sum_{k \in \mathcal{K}} \widetilde{c}_k \left(\tilde{I}_t^{(k)}, d_t^{(k)} \right) \right)^2 \right] \\
&= \mathbb{E}_t \left[\sum_{j \in \mathcal{K}} \frac{1}{\mathbb{P}_t \left(I_t^{(j)} \geq \tilde{I}_t^{(j)} \right)^2} \mathbb{I}_{\left\{ I_t^{(j)} \geq \tilde{I}_t^{(j)} \right\}} \ell_j \left(\tilde{I}_t^{(j)}, d_t^{(j)} \right) \sum_{k \in \mathcal{K}} \mathbb{I}_{\left\{ I_t^{(k)} \geq \tilde{I}_t^{(k)} \right\}} \ell_k \left(\tilde{I}_t^{(k)}, d_t^{(k)} \right) \right] \\
&\leq 2K\beta \mathbb{E}_t \left[\sum_{j \in \mathcal{K}} \frac{1}{\mathbb{P}_t \left(I_t^{(j)} \geq \tilde{I}_t^{(j)} \right)^2} \mathbb{I}_{\left\{ I_t^{(j)} \geq \tilde{I}_t^{(j)} \right\}} \ell_j \left(\tilde{I}_t^{(j)}, d_t^{(j)} \right) \right] \\
&= 2K\beta \mathbb{E}_t \left[\sum_{j \in \mathcal{K}} \frac{\ell_j \left(\tilde{I}_t^{(j)}, d_t^{(j)} \right)}{\mathbb{P}_t \left(I_t^{(j)} \geq \tilde{I}_t^{(j)} \right)} \right] \\
&\leq 4K\beta^2 \sum_{j=1}^{K} \mathbb{E}_t \left[\sum_{i=1}^{N} \frac{\mathbb{P}_t \left(I_t^{(j)} = i \right)}{\mathbb{P}_t \left(I_t^{(j)} \geq i \right)} \right],
\end{aligned}
$$

where the inequalities follow from bounding from above $\ell_j(\cdot,\cdot)$ by 2β.

It remains to bound the sums within the expectation, for all j. To this end, we observe that our exploration distribution μ guarantees that $\mathbb{P}\left[I_t^{(j)} = i \right] \geq \frac{\gamma}{NK}$ holds for all i, j. Given this lower bound, we

can apply Lemma 4 with $\omega = \frac{\gamma}{NK}$ as done in the proof of Theorem 1:

$$\sum_{i=1}^{N} \frac{\mathbb{P}_t\left(I_t^{(j)} = i\right)}{\mathbb{P}_t\left(I_t^{(j)} \geq i\right)} \leq 5\log\left(\frac{3KN}{\gamma} + 3\right).$$

Putting everything together, we obtain the desired result. ■

Proof of Theorem 3

By following closely the proof of Theorem 1 with our definition of μ and applying Eq. (8), we have that

$$\begin{aligned}
\mathbb{E}[\mathcal{R}_c(T)] \leq & \frac{\log|\mathcal{A}_r|}{\eta} + \gamma \sum_{t\in\mathcal{T}} \sum_{(i^{(1)},\ldots,i^{(K)})\in\mathcal{A}_r} \mu_{(i^{(1)},\ldots,i^{(K)})} \mathbb{E}_t\left[\widetilde{c}\left(i^{(1)},\ldots,i^{(K)}, r, d_t^{(1)},\ldots,d_t^{(K)}\right)\right] \\
& + \frac{\eta}{2} \sum_{t\in\mathcal{T}} \sum_{(i^{(1)},\ldots,i^{(K)})\in\mathcal{A}_r} p_{(i^{(1)},\ldots,i^{(K)})}(t) \mathbb{E}_t\left[\widetilde{c}\left(i^{(1)},\ldots,i^{(K)}, r, d_t^{(1)},\ldots,d_t^{(K)}\right)^2\right] \\
\leq & \frac{K\log N}{\eta} + \gamma\big(f + K(2\beta + f)\big)T + \eta T\left(10K^2\beta^2\log\left(\frac{3KN}{\gamma} + 3\right) + (f + K\beta)^2\right),
\end{aligned}$$

where the last step uses Eq. (9) and Lemma 5. Substituting the prescribed values for γ and η yields the statement of the theorem.