A Proofs of Main Theorems

A.1 Proof of Lemma 1

Let $R_t = R(A_t, w_t)$ be the stochastic regret of `CombUCB1` at time t, where A_t and w_t are the solution and the weights of the items at time t, respectively. Furthermore, let $\mathcal{E}_t = \left\{\exists e \in E : \left|\bar{w}(e) - \hat{w}_{T_{t-1}(e)}(e)\right| \geq c_{t-1,T_{t-1}(e)}\right\}$ be the event that $\bar{w}(e)$ is outside of the high-probability confidence interval around $\hat{w}_{T_{t-1}(e)}(e)$ for some item e at time t; and let $\overline{\mathcal{E}}_t$ be the complement of $\mathcal{E}_t$, $\bar{w}(e)$ is in the high-probability confidence interval around $\hat{w}_{T_{t-1}(e)}(e)$ for all e at time t. Then we can decompose the regret of `CombUCB1` as:

$$R(n) = \mathbb{E}\left[\sum_{t=1}^{t_0-1} R_t\right] + \mathbb{E}\left[\sum_{t=t_0}^{n} \mathbb{1}\{\mathcal{E}_t\} R_t\right] + \mathbb{E}\left[\sum_{t=t_0}^{n} \mathbb{1}\{\overline{\mathcal{E}}_t\} R_t\right].$$

Now we bound each term in our regret decomposition.

The regret of the initialization, $\mathbb{E}\left[\sum_{t=1}^{t_0-1} R_t\right]$, is bounded by KL because Algorithm 2 terminates in at most L steps, and $R_t \leq K$ for any A_t and w_t.

The second term in our regret decomposition, $\mathbb{E}\left[\sum_{t=t_0}^{n} \mathbb{1}\{\mathcal{E}_t\} R_t\right]$, is small because all of our confidence intervals hold with high probability. In particular, for any e, s, and t:

$$P(|\bar{w}(e) - \hat{w}_s(e)| \geq c_{t,s}) \leq 2\exp[-3\log t],$$

and therefore:

$$\mathbb{E}\left[\sum_{t=t_0}^{n} \mathbb{1}\{\mathcal{E}_t\}\right] \leq \sum_{e \in E}\sum_{t=1}^{n}\sum_{s=1}^{t} P(|\bar{w}(e) - \hat{w}_s(e)| \geq c_{t,s}) \leq 2\sum_{e \in E}\sum_{t=1}^{n}\sum_{s=1}^{t}\exp[-3\log t] \leq 2\sum_{e \in E}\sum_{t=1}^{n} t^{-2} \leq \frac{\pi^2}{3}L.$$

Since $R_t \leq K$ for any A_t and w_t, $\mathbb{E}\left[\sum_{t=t_0}^{n} \mathbb{1}\{\mathcal{E}_t\} R_t\right] \leq \frac{\pi^2}{3}KL$.

Finally, we rewrite the last term in our regret decomposition as:

$$\mathbb{E}\left[\sum_{t=t_0}^{n} \mathbb{1}\{\overline{\mathcal{E}}_t\} R_t\right] \overset{\text{(a)}}{=} \sum_{t=t_0}^{n} \mathbb{E}\left[\mathbb{1}\{\overline{\mathcal{E}}_t\}\mathbb{E}\left[R_t \mid A_t\right]\right] \overset{\text{(b)}}{=} \mathbb{E}\left[\sum_{t=t_0}^{n} \Delta_{A_t}\mathbb{1}\{\overline{\mathcal{E}}_t, \Delta_{A_t} > 0\}\right].$$

In equality (a), the outer expectation is over the history of the agent up to time t, which in turn determines A_t and $\overline{\mathcal{E}}_t$; and $\mathbb{E}\left[R_t \mid A_t\right]$ is the expected regret at time t conditioned on solution A_t. Equality (b) follows from $\Delta_{A_t} = \mathbb{E}\left[R_t \mid A_t\right]$. Now we bound $\Delta_{A_t}\mathbb{1}\left\{\overline{\mathcal{E}}_t, \Delta_{A_t} > 0\right\}$ for any suboptimal A_t. The bound is derived based on two facts. First, when `CombUCB1` chooses A_t, $f(A_t, U_t) \geq f(A^*, U_t)$. This further implies that $\sum_{e \in A_t \setminus A^*} U_t(e) \geq \sum_{e \in A^* \setminus A_t} U_t(e)$. Second, when event $\overline{\mathcal{E}}_t$ happens, $\left|\bar{w}(e) - \hat{w}_{T_{t-1}(e)}(e)\right| < c_{t-1,T_{t-1}(e)}$ for all items e. Therefore:

$$\sum_{e \in A_t \setminus A^*} \bar{w}(e) + 2\sum_{e \in A_t \setminus A^*} c_{t-1,T_{t-1}(e)} \geq \sum_{e \in A_t \setminus A^*} U_t(e) \geq \sum_{e \in A^* \setminus A_t} U_t(e) \geq \sum_{e \in A^* \setminus A_t} \bar{w}(e),$$

and $2\sum_{e \in A_t \setminus A^*} c_{t-1,T_{t-1}(e)} \geq \Delta_{A_t}$ follows from the observation that $\Delta_{A_t} = \sum_{e \in A^* \setminus A_t} \bar{w}(e) - \sum_{e \in A_t \setminus A^*} \bar{w}(e)$. Now note that $c_{n,T_{t-1}(e)} \geq c_{t-1,T_{t-1}(e)}$ for any time $t \leq n$. Therefore, the event $\mathcal{F}_t$ in (3) must happen and:

$$\mathbb{E}\left[\sum_{t=t_0}^{n} \Delta_{A_t}\mathbb{1}\{\overline{\mathcal{E}}_t, \Delta_{A_t} > 0\}\right] \leq \mathbb{E}\left[\sum_{t=t_0}^{n} \Delta_{A_t}\mathbb{1}\{\mathcal{F}_t\}\right].$$

This concludes our proof.

A.2 Proof of Theorem 2

By Lemma 1, it remains to bound $\hat{R}(n) = \sum_{t=t_0}^{n} \Delta_{A_t}\mathbb{1}\{\mathcal{F}_t\}$, where the event $\mathcal{F}_t$ is defined in (3). By Lemma 2 and from the assumption that $\Delta_{A_t} = \Delta$ for all suboptimal A_t, it follows that:

$$\hat{R}(n) = \Delta\sum_{t=t_0}^{n} \mathbb{1}\{\mathcal{F}_t\} = \Delta\sum_{t=t_0}^{n} \mathbb{1}\{G_{1,t}, \Delta_{A_t} > 0\} + \Delta\sum_{t=t_0}^{n} \mathbb{1}\{G_{2,t}, \Delta_{A_t} > 0\}.$$

To bound the above quantity, it is sufficient to bound the number of times that events $G_{1,t}$ and $G_{2,t}$ happen. Then we set the tunable parameters d and α such that the two counts are of the same magnitude.

Claim 1. *Event $G_{1,t}$ happens at most $\frac{\alpha}{d} K^2 L \frac{6}{\Delta^2} \log n$ times.*

Proof. Recall that event $G_{1,t}$ can happen only if at least d chosen suboptimal items are not observed "sufficiently often" up to time t, $T_{t-1}(e) \leq \alpha K^2 \frac{6}{\Delta^2} \log n$ for at least d items in $\tilde{A}_t$. After the event happens, the observation counters of these items increase by one. Therefore, after the event happens $\frac{\alpha}{d} K^2 L \frac{6}{\Delta^2} \log n$ times, all suboptimal items are guaranteed to be observed at least $\alpha K^2 \frac{6}{\Delta^2} \log n$ times and $G_{1,t}$ cannot happen anymore. ■

Claim 2. *Event $G_{2,t}$ happens at most $\frac{\alpha d^2}{(\sqrt{\alpha}-1)^2} L \frac{6}{\Delta^2} \log n$ times.*

Proof. Event $G_{2,t}$ can happen only if there exists $e \in \tilde{A}_t$ such that $T_{t-1}(e) \leq \frac{\alpha d^2}{(\sqrt{\alpha}-1)^2} \frac{6}{\Delta^2} \log n$. After the event happens, the observation counter of item e increases by one. Therefore, the number of times that event $G_{2,t}$ can happen is bounded trivially by $\frac{\alpha d^2}{(\sqrt{\alpha}-1)^2} L \frac{6}{\Delta^2} \log n$. ■

Based on Claims 1 and 2, $\hat{R}(n)$ is bounded as:

$$\hat{R}(n) \leq \left(\frac{\alpha}{d} K^2 + \frac{\alpha d^2}{(\sqrt{\alpha}-1)^2} \right) L \frac{6}{\Delta} \log n \,.$$

Finally, we choose $\alpha = 4$ and $d = K^{\frac{2}{3}}$; and it follows that the regret is bounded as:

$$R(n) \leq \mathbb{E}\left[\hat{R}(n)\right] + \left(\frac{\pi^2}{3} + 1\right) KL \leq K^{\frac{4}{3}} L \frac{48}{\Delta} \log n + \left(\frac{\pi^2}{3} + 1\right) KL \,.$$

A.3 Proof of Theorem 3

Let $\mathcal{F}_t$ be the event in (3). By Lemmas 1 and 2, it remains to bound:

$$\hat{R}(n) = \sum_{t=t_0}^{n} \Delta_{A_t} \mathbb{1}\{\mathcal{F}_t\} = \sum_{t=t_0}^{n} \Delta_{A_t} \mathbb{1}\{G_{1,t}, \Delta_{A_t} > 0\} + \sum_{t=t_0}^{n} \Delta_{A_t} \mathbb{1}\{G_{2,t}, \Delta_{A_t} > 0\} \,.$$

In the next step, we introduce item-specific variants of events $G_{1,t}$ (6) and $G_{2,t}$ (7), and then associate the regret at time t with these events. In particular, let:

$$G_{e,1,t} = G_{1,t} \cap \left\{ e \in \tilde{A}_t, T_{t-1}(e) \leq \alpha K^2 \frac{6}{\Delta_{A_t}^2} \log n \right\} \tag{14}$$

$$G_{e,2,t} = G_{2,t} \cap \left\{ e \in \tilde{A}_t, T_{t-1}(e) \leq \frac{\alpha d^2}{(\sqrt{\alpha}-1)^2} \frac{6}{\Delta_{A_t}^2} \log n \right\} \tag{15}$$

be the events that item e is not observed "sufficiently often" under events $G_{1,t}$ and $G_{2,t}$, respectively. Then by the definitions of the above events, it follows that:

$$\mathbb{1}\{G_{1,t}, \Delta_{A_t} > 0\} \leq \frac{1}{d} \sum_{e \in \tilde{E}} \mathbb{1}\{G_{e,1,t}, \Delta_{A_t} > 0\}$$
$$\mathbb{1}\{G_{2,t}, \Delta_{A_t} > 0\} \leq \sum_{e \in \tilde{E}} \mathbb{1}\{G_{e,2,t}, \Delta_{A_t} > 0\} \,,$$

where $\tilde{E} = E \setminus A^*$ is the set of subptimal items; and we bound $\hat{R}(n)$ as:

$$\hat{R}(n) \leq \sum_{e \in \tilde{E}} \sum_{t=t_0}^{n} \mathbb{1}\{G_{e,1,t}, \Delta_{A_t} > 0\} \frac{\Delta_{A_t}}{d} + \sum_{e \in \tilde{E}} \sum_{t=t_0}^{n} \mathbb{1}\{G_{e,2,t}, \Delta_{A_t} > 0\} \Delta_{A_t} \,.$$

Let each item e be contained in N_e suboptimal solutions and $\Delta_{e,1} \geq \ldots \geq \Delta_{e,N_e}$ be the gaps of these solutions, ordered from the largest gap to the smallest one. Then $\hat{R}(n)$ can be further bounded as:

$$\begin{aligned}
\hat{R}(n) &\leq \sum_{e \in \tilde{E}} \sum_{t=t_0}^{n} \sum_{k=1}^{N_e} \mathbb{1}\{G_{e,1,t}, \Delta_{A_t} = \Delta_{e,k}\} \frac{\Delta_{e,k}}{d} + \sum_{e \in \tilde{E}} \sum_{t=t_0}^{n} \sum_{k=1}^{N_e} \mathbb{1}\{G_{e,2,t}, \Delta_{A_t} = \Delta_{e,k}\} \Delta_{e,k} \\
&\overset{\text{(a)}}{\leq} \sum_{e \in \tilde{E}} \sum_{t=t_0}^{n} \sum_{k=1}^{N_e} \mathbb{1}\left\{e \in \tilde{A}_t, T_{t-1}(e) \leq \alpha K^2 \frac{6}{\Delta_{e,k}^2} \log n, \Delta_{A_t} = \Delta_{e,k}\right\} \frac{\Delta_{e,k}}{d} + \\
&\quad \sum_{e \in \tilde{E}} \sum_{t=t_0}^{n} \sum_{k=1}^{N_e} \mathbb{1}\left\{e \in \tilde{A}_t, T_{t-1}(e) \leq \frac{\alpha d^2}{(\sqrt{\alpha}-1)^2} \frac{6}{\Delta_{e,k}^2} \log n, \Delta_{A_t} = \Delta_{e,k}\right\} \Delta_{e,k} \\
&\overset{\text{(b)}}{\leq} \sum_{e \in \tilde{E}} \frac{6\alpha K^2 \log n}{d} \left[\Delta_{e,1} \frac{1}{\Delta_{e,1}^2} + \sum_{k=2}^{N_e} \Delta_{e,k} \left(\frac{1}{\Delta_{e,k}^2} - \frac{1}{\Delta_{e,k-1}^2}\right)\right] + \\
&\quad \sum_{e \in \tilde{E}} \frac{6\alpha d^2 \log n}{(\sqrt{\alpha}-1)^2} \left[\Delta_{e,1} \frac{1}{\Delta_{e,1}^2} + \sum_{k=2}^{N_e} \Delta_{e,k} \left(\frac{1}{\Delta_{e,k}^2} - \frac{1}{\Delta_{e,k-1}^2}\right)\right] \\
&\overset{\text{(c)}}{<} \sum_{e \in \tilde{E}} \left(\frac{\alpha}{d} K^2 + \frac{\alpha d^2}{(\sqrt{\alpha}-1)^2}\right) \frac{12}{\Delta_{e,\min}} \log n \,,
\end{aligned}$$

where inequality (a) is by the definitions of events $G_{e,1,t}$ and $G_{e,2,t}$, inequality (b) is from the solution to:

$$\max_{A_1,\ldots,A_n} \sum_{t=t_0}^{n} \sum_{k=1}^{N_e} \mathbb{1}\left\{e \in \tilde{A}_t, T_{t-1}(e) \leq \frac{C}{\Delta_{e,k}^2} \log n, \Delta_{A_t} = \Delta_{e,k}\right\} \Delta_{e,k}$$

for appropriate C, and inequality (c) follows from Lemma 3 of Kveton *et al.* [12]:

$$\left[\Delta_{e,1} \frac{1}{\Delta_{e,1}^2} + \sum_{k=2}^{N_e} \Delta_{e,k} \left(\frac{1}{\Delta_{e,k}^2} - \frac{1}{\Delta_{e,k-1}^2}\right)\right] < \frac{2}{\Delta_{e,N_e}} = \frac{2}{\Delta_{e,\min}} \,. \tag{16}$$

Finally, we choose $\alpha = 4$ and $d = K^{\frac{2}{3}}$; and it follows that the regret is bounded as:

$$R(n) \leq \mathbb{E}\left[\hat{R}(n)\right] + \left(\frac{\pi^2}{3} + 1\right) KL \leq \sum_{e \in \tilde{E}} K^{\frac{4}{3}} \frac{96}{\Delta_{e,\min}} \log n + \left(\frac{\pi^2}{3} + 1\right) KL \,.$$

A.4 Proof of Theorem 4

The first step of the proof is identical to that of Theorem 2. By Lemma 1, it remains to bound $\hat{R}(n) = \sum_{t=t_0}^{n} \Delta_{A_t} \mathbb{1}\{\mathcal{F}_t\}$, where the event $\mathcal{F}_t$ is defined in (3). By Lemma 3 and from the assumption that $\Delta_{A_t} = \Delta$ for all suboptimal A_t, it follows that:

$$\hat{R}(n) = \Delta \sum_{t=t_0}^{n} \mathbb{1}\{\mathcal{F}_t\} = \Delta \sum_{i=1}^{\infty} \sum_{t=t_0}^{n} \mathbb{1}\{G_{i,t}, \Delta_{A_t} > 0\} \,.$$

Note that $\Delta_{A_t} > 0$ implies $\Delta_{A_t} = \Delta$. Therefore, $m_{i,t}$ does not depend on t and we denote it by $m_i = \alpha_i \frac{K^2}{\Delta^2} \log n$. Based on the same argument as in Claim 1, event $G_{i,t}$ cannot happen more than $\frac{L m_i}{\beta_i K}$ times, because at least $\beta_i K$ items that are observed at most m_i times have their observation counters incremented in each event $G_{i,t}$. Therefore:

$$\hat{R}(n) \leq \Delta \sum_{i=1}^{\infty} \frac{L m_i}{\beta_i K} = KL \frac{1}{\Delta} \left[\sum_{i=1}^{\infty} \frac{\alpha_i}{\beta_i}\right] \log n \,. \tag{17}$$

It remains to choose (α_i) and (β_i) such that:

- $\lim_{i \to \infty} \alpha_i = \lim_{i \to \infty} \beta_i = 0$;
- Monotonicity conditions in (9) and (10) hold;
- Inequality (12) holds, $\sqrt{6} \sum_{i=1}^{\infty} \frac{\beta_{i-1} - \beta_i}{\sqrt{\alpha_i}} \leq 1$;
- $\sum_{i=1}^{\infty} \frac{\alpha_i}{\beta_i}$ is minimized.

We choose (α_i) and (β_i) to be geometric sequences, $\beta_i = \beta^i$ and $\alpha_i = d\alpha^i$ for $0 < \alpha, \beta < 1$ and $d > 0$. For this setting, $\alpha_i \to 0$ and $\beta_i \to 0$, and the monotonicity conditions are also satisfied. Moreover, if $\beta < \sqrt{\alpha}$, we have:

$$\sqrt{6} \sum_{i=1}^{\infty} \frac{\beta_{i-1} - \beta_i}{\sqrt{\alpha_i}} = \sqrt{6} \sum_{i=1}^{\infty} \frac{\beta^{i-1} - \beta^i}{\sqrt{d\alpha^i}} = \sqrt{\frac{6}{d}} \frac{1 - \beta}{\sqrt{\alpha} - \beta} \leq 1$$

provided that $d \geq 6 \left(\frac{1-\beta}{\sqrt{\alpha} - \beta}\right)^2$. Furthermore, if $\alpha < \beta$, we have:

$$\sum_{i=1}^{\infty} \frac{\alpha_i}{\beta_i} = \sum_{i=1}^{\infty} \frac{d\alpha^i}{\beta^i} = \frac{d\alpha}{\beta - \alpha} .$$

Given the above, the best choice of d is $6 \left(\frac{1-\beta}{\sqrt{\alpha} - \beta}\right)^2$ and the problem of minimizing the constant in our regret bound can be written as:

$$\begin{aligned} \inf_{\alpha, \beta} \quad & 6 \left(\frac{1 - \beta}{\sqrt{\alpha} - \beta}\right)^2 \frac{\alpha}{\beta - \alpha} \\ \text{s.t.} \quad & 0 < \alpha < \beta < \sqrt{\alpha} < 1 . \end{aligned}$$

We find the solution to the above problem numerically, and determine it to be $\alpha = 0.1459$ and $\beta = 0.2360$. For these α and β, $6 \left(\frac{1-\beta}{\sqrt{\alpha} - \beta}\right)^2 \frac{\alpha}{\beta - \alpha} < 267$. We apply this upper bound to (17) and it follows that the regret is bounded as:

$$R(n) \leq \mathbb{E}\left[\hat{R}(n)\right] + \left(\frac{\pi^2}{3} + 1\right) KL \leq KL \frac{267}{\Delta} \log n + \left(\frac{\pi^2}{3} + 1\right) KL .$$

A.5 Proof of Theorem 5

Let $\mathcal{F}_t$ be the event in (3). By Lemmas 1 and 3, it remains to bound:

$$\hat{R}(n) = \sum_{t=t_0}^{n} \Delta_{A_t} \mathbb{1}\{\mathcal{F}_t\} = \sum_{i=1}^{\infty} \sum_{t=t_0}^{n} \Delta_{A_t} \mathbb{1}\{G_{i,t}, \Delta_{A_t} > 0\} .$$

In the next step, we define item-specific variants of events $G_{i,t}$ (11) and associate the regret at time t with these events. In particular, let:

$$G_{e,i,t} = G_{i,t} \cap \left\{e \in \tilde{A}_t, T_{t-1}(e) \leq m_{i,t}\right\} \tag{18}$$

be the event that item e is not observed “sufficiently often” under event $G_{i,t}$. Then it follows that:

$$\mathbb{1}\{G_{i,t}, \Delta_{A_t} > 0\} \leq \frac{1}{\beta_i K} \sum_{e \in \tilde{E}} \mathbb{1}\{G_{e,i,t}, \Delta_{A_t} > 0\} ,$$

because at least $\beta_i K$ items are not observed “sufficiently often” under event $G_{i,t}$. Therefore, we can bound $\hat{R}(n)$ as:

$$\hat{R}(n) \leq \sum_{e \in \tilde{E}} \sum_{i=1}^{\infty} \sum_{t=t_0}^{n} \mathbb{1}\{G_{e,i,t}, \Delta_{A_t} > 0\} \frac{\Delta_{A_t}}{\beta_i K} .$$

Let each item e be contained in N_e suboptimal solutions and $\Delta_{e,1} \geq \ldots \geq \Delta_{e,N_e}$ be the gaps of these solutions, ordered from the largest gap to the smallest one. Then $\hat{R}(n)$ can be further bounded as:

$$\begin{aligned}
\hat{R}(n) &\leq \sum_{e \in \tilde{E}} \sum_{i=1}^{\infty} \sum_{t=t_0}^{n} \sum_{k=1}^{N_e} \mathbb{1}\{G_{e,i,t}, \Delta_{A_t} = \Delta_{e,k}\} \frac{\Delta_{e,k}}{\beta_i K} \\
&\overset{\text{(a)}}{\leq} \sum_{e \in \tilde{E}} \sum_{i=1}^{\infty} \sum_{t=t_0}^{n} \sum_{k=1}^{N_e} \mathbb{1}\left\{e \in \tilde{A}_t, T_{t-1}(e) \leq \alpha_i \frac{K^2}{\Delta_{e,k}^2} \log n, \Delta_{A_t} = \Delta_{e,k}\right\} \frac{\Delta_{e,k}}{\beta_i K} \\
&\overset{\text{(b)}}{\leq} \sum_{e \in \tilde{E}} \sum_{i=1}^{\infty} \frac{\alpha_i K \log n}{\beta_i} \left[\Delta_{e,1} \frac{1}{\Delta_{e,1}^2} + \sum_{k=2}^{N_e} \Delta_{e,k} \left(\frac{1}{\Delta_{e,k}^2} - \frac{1}{\Delta_{e,k-1}^2}\right)\right] \\
&\overset{\text{(c)}}{<} \sum_{e \in \tilde{E}} \sum_{i=1}^{\infty} \frac{\alpha_i K \log n}{\beta_i} \frac{2}{\Delta_{e,\min}} \\
&= \sum_{e \in \tilde{E}} K \frac{2}{\Delta_{e,\min}} \left[\sum_{i=1}^{\infty} \frac{\alpha_i}{\beta_i}\right] \log n\,,
\end{aligned}$$

where inequality (a) is by the definition of event $G_{e,i,t}$, inequality (b) follows from the solution to:

$$\max_{A_1,\ldots,A_n} \sum_{t=t_0}^{n} \sum_{k=1}^{N_e} \mathbb{1}\left\{e \in \tilde{A}_t, T_{t-1}(e) \leq \alpha_i \frac{K^2}{\Delta_{e,k}^2} \log n, \Delta_{A_t} = \Delta_{e,k}\right\} \frac{\Delta_{e,k}}{\beta_i K}\,,$$

and inequality (c) follows from (16). For the same (α_i) and (β_i) as in Theorem 4, we have $\sum_{i=1}^{\infty} \frac{\alpha_i}{\beta_i} < 267$ and it follows that the regret is bounded as:

$$R(n) \leq \mathbb{E}\left[\hat{R}(n)\right] + \left(\frac{\pi^2}{3} + 1\right) KL \leq \sum_{e \in \tilde{E}} K \frac{534}{\Delta_{e,\min}} \log n + \left(\frac{\pi^2}{3} + 1\right) KL\,.$$

A.6 Proof of Theorem 6

The key idea is to decompose the regret of `CombUCB1` into two parts, where the gaps are larger than ϵ and at most ϵ. We analyze each part separately and then set ϵ to get the desired result.

By Lemma 1, it remains to bound $\hat{R}(n) = \sum_{t=t_0}^{n} \Delta_{A_t} \mathbb{1}\{\mathcal{F}_t\}$, where the event $\mathcal{F}_t$ is defined in (3). We partition $\hat{R}(n)$ as:

$$\begin{aligned}
\hat{R}(n) &= \sum_{t=t_0}^{n} \Delta_{A_t} \mathbb{1}\{\mathcal{F}_t, \Delta_{A_t} < \epsilon\} + \sum_{t=t_0}^{n} \Delta_{A_t} \mathbb{1}\{\mathcal{F}_t, \Delta_{A_t} \geq \epsilon\} \\
&\leq \epsilon n + \sum_{t=t_0}^{n} \Delta_{A_t} \mathbb{1}\{\mathcal{F}_t, \Delta_{A_t} \geq \epsilon\}
\end{aligned}$$

and bound the first term trivially. The second term is bounded in the same way as $\hat{R}(n)$ in the proof of Theorem 5, except that we only consider the gaps $\Delta_{e,k} \geq \epsilon$. Therefore, $\Delta_{e,\min} \geq \epsilon$ and we get:

$$\sum_{t=t_0}^{n} \Delta_{A_t} \mathbb{1}\{\mathcal{F}_t, \Delta_{A_t} \geq \epsilon\} \leq \sum_{e \in \tilde{E}} K \frac{534}{\epsilon} \log n \leq KL \frac{534}{\epsilon} \log n\,.$$

Based on the above inequalities:

$$R(n) \leq \frac{534KL}{\epsilon} \log n + \epsilon n + \left(\frac{\pi^2}{3} + 1\right) KL\,.$$

Finally, we choose $\epsilon = \sqrt{\frac{534KL \log n}{n}}$ and get:

$$R(n) \leq 2\sqrt{534KLn \log n} + \left(\frac{\pi^2}{3} + 1\right) KL < 47\sqrt{KLn \log n} + \left(\frac{\pi^2}{3} + 1\right) KL\,,$$

which concludes our proof.

B Technical Lemmas

Lemma 4. *Let S_i, $\bar{S}_i$, and m_i be defined as in Lemma 3; and $|S_i| < \beta_i K$ for all $i > 0$. Then:*

$$\sum_{i=1}^{\infty} \frac{|\bar{S}_i \setminus \bar{S}_{i-1}|}{\sqrt{m_i}} < \sum_{i=1}^{\infty} \frac{(\beta_{i-1} - \beta_i)K}{\sqrt{m_i}} \,.$$

Proof. The lemma is proved as:

$$\begin{aligned}
\sum_{i=1}^{\infty} |\bar{S}_i \setminus \bar{S}_{i-1}| \frac{1}{\sqrt{m_i}} &= \sum_{i=1}^{\infty} (|S_{i-1} \setminus S_i|) \frac{1}{\sqrt{m_i}} \\
&= \sum_{i=1}^{\infty} (|S_{i-1}| - |S_i|) \frac{1}{\sqrt{m_i}} \\
&= \frac{|S_0|}{\sqrt{m_1}} + \sum_{i=1}^{\infty} |S_i| \left(\frac{1}{\sqrt{m_{i+1}}} - \frac{1}{\sqrt{m_i}} \right) \\
&< \frac{\beta_0 K}{\sqrt{m_1}} + \sum_{i=1}^{\infty} \beta_i K \left(\frac{1}{\sqrt{m_{i+1}}} - \frac{1}{\sqrt{m_i}} \right) \\
&= \sum_{i=1}^{\infty} (\beta_{i-1} - \beta_i) K \frac{1}{\sqrt{m_i}} \,.
\end{aligned}$$

The first two equalities follow from the definitions of $\bar{S}_i$ and S_i. The inequality follows from the facts that $|S_i| < \beta_i K$ for all $i > 0$ and $|S_0| \leq \beta_0 K$. ■