1 Proof of Theorem 1

Theorem 1. *Define* $m = min\left\{1 \leq i \leq N : q_{i+1} \geq \frac{1}{i}\right\}$ *Then Algorithm 1 satisfies*

$$R(T) \in \mathcal{O}\left(T^{2/3}m^{1/3}log(KT)^{1/3}\right) .$$

Let $A = \{(i, j) : i \leq m, j = 1\}$ be the set of infrequently observed arms

For the frequently observed arms, $(i, j) \notin A$ we have:

$$\hat{\mu}_{i,j} = \frac{2}{h}\frac{\sum_{t=1}^{h/2} \mathbb{1}\{X_{i,t} = j\}\, r_t}{q_i^j(1 - q_i)^{1-j}} \tag{1}$$

Let $Z_{t,ij} = \mathbb{1}\{Y_t = 1, X_{t,i} = j\} \sim Bernoulli(q_{ij}\mu_{ij})$,

Chernoff's inequality gives

$$P(\hat{\mu}_i - \mu_i > \frac{D}{2}) \leq e^{-hD^2/24m} \tag{2}$$

The algorithm explicitly plays each of the infrequently observed arms, $(i, j) \in A$, $\frac{h}{2m}$ times. So:

$$P(\hat{\mu}_i - \mu_i > \frac{D}{2}) \leq e^{-hD^2/4m} \tag{3}$$

So for all the arms

$$P(\hat{\mu}_i - \mu_i > \frac{D}{2}) \leq e^{-hD^2/24m} \tag{4}$$

and

$$P(\Delta_{\hat{i^*}} > D) \leq Ke^{-hD^2/24m} \tag{5}$$

If we let $D = \sqrt{\frac{24m\log(hK)}{h}}$

$$\begin{aligned} R_T &\leq h + T\left(\sqrt{\frac{24m\log(hK)}{h}} + \frac{1}{h}\right) && (6)\\ &\leq h + T\left(\sqrt{\frac{24m\log(TK)}{h}} + \frac{1}{h}\right) && (7)\end{aligned}$$

Let $h = T^{2/3}m^{1/3}\log(TK)^{1/3}$

$$\begin{aligned} R_T &\leq 6T^{2/3}m^{1/3}log(KT)^{1/3} + T^{1/3}m^{-1/3}log(KT)^{-1/3} && (8)\\ &\leq 7T^{2/3}m^{1/3}log(KT)^{1/3} && (9)\end{aligned}$$

2 Proof of Theorem 2

Theorem 2. *Define* $m = \min\{2 \leq i \leq N : q_i \geq 1/i\}$. *Then Algorithm 2 satisfies*

$$R^{simple}(T) \in O\left(\sqrt{\frac{m}{T}\log\left(\frac{NT}{m}\right)}\right) .$$

Lemma 3. $\mathrm{P}\left(|\hat{q}_i - q_i| \geq \sqrt{\frac{6q_i}{T}\log\frac{2}{\delta}}\right) \leq \delta.$

Proof. Let $Z_t = \mathbb{1}\{X_{i,t} = 1\} \in \{0, 1\}$. Then

$$\hat{q}_i = \frac{2}{T}\sum_{t=1}^{T/2} Z_t .$$

Now $Z_1, \ldots, Z_{T/2}$ is an i.i.d. sequence of Bernoulli random variables with with mean q_i. The result follows from the Chernoff bound. □

Lemma 4. *Let* $\delta > 0$. *If* $h \geq 24m\log\frac{4N}{\delta}$

then

$$\mathrm{P}\left(\hat{m} < \frac{2}{3}m\right) \leq \delta \;\; \textit{and} \;\; \mathrm{P}\left(\hat{m} > 2m\right) \leq \delta .$$

Proof. Let $\boldsymbol{q}^b$ and $\boldsymbol{q}^{ub}$ and be the maximally balanced and unbalanced $\boldsymbol{q}$ for a given m respectively.

$$q_i^{ub} = \begin{cases} 0 & \text{if } i \leq m \\ \frac{1}{m} & \text{otherwise} . \end{cases}$$

$$q_i^{b} < \begin{cases} \frac{1}{m} & \text{if } i \leq m \\ 1 & \text{otherwise} . \end{cases}$$

For $\hat{m}$ to over-estimate m, we must identify some balanced arms as unbalanced. For $\hat{m}$ to under-estimate m, we must identify some unbalanced arms as balanced.

$$\mathrm{P}\left(\hat{m} > 2m\right) \leq \mathrm{P}\left(\hat{m} > 2m | \boldsymbol{q} = \boldsymbol{q}^{ub}\right)$$

$$\mathrm{P}\left(\hat{m} < \frac{2}{3}m\right) \leq \mathrm{P}\left(\hat{m} < \frac{2}{3}m | \boldsymbol{q} = \boldsymbol{q}^{b}\right)$$

Given $\boldsymbol{q} = \boldsymbol{q}^{ub}$, we have by Lemma 3, with probability at least $1 - \delta$ that:

$$\begin{aligned}
&|\hat{q}_i - q_i| \leq \begin{cases} 0 & \text{if } i \leq m \\ \sqrt{\frac{6}{mh}\log\frac{2}{\delta}} & \text{otherwise} \end{cases} \\
\implies & \begin{cases} (\forall i \leq m) & |\hat{q}_i - 0| = 0 \\ (\forall i > m) & \left|\hat{q}_i - \frac{1}{m}\right| \leq \frac{1}{2m} \end{cases} \text{, taking the union bound and assuming } h \geq 24m\log\tfrac{4N}{\delta} \\
\implies & \begin{cases} (\forall i \leq m) & \hat{q}_i = 0 \\ (\forall i > m) & \hat{q}_i \in [\frac{1}{2m}, \frac{3}{2m}] \end{cases} \\
\implies & \hat{m} \leq 2m
\end{aligned}$$

Given $\boldsymbol{q} = \boldsymbol{q}^b$, we have by Lemma 3, with probability at least $1 - \delta$ that:

$$
\begin{aligned}
&|\hat{q}_i - q_i| \leq \sqrt{\frac{6}{mh} \log \frac{2}{\delta}} \qquad \text{if } i \leq m \\
\implies &(\forall i \leq m) \qquad \hat{q}_i \leq \frac{3}{2m} \\
\implies &\hat{m} \geq \frac{2m}{3}
\end{aligned}
$$

□

Proof of Theorem 2. for $(i, j) \in A$, the algorithm explicitly selects the action, $X_i = j$, $\frac{h}{2\hat{m}}$ times.

$$
\hat{\mu}_{i,j} = \frac{2\hat{m}}{h} \sum_{t=1}^{h/2\hat{m}} r_t(X_i = j)
$$

Via Hoeffding's Inequality

$$
\mathrm{P}\left(|\hat{\mu}_{i,j} - \mu_{i,j}| > \epsilon\right) \leq 2 \exp - \frac{h\epsilon^2}{\hat{m}}
$$

for $(i, j) \notin A$, the algorithm has observed the reward given $X_i = j$ at least $\frac{h}{2\hat{m}}$ times.

$$
\begin{aligned}
(i, j) \notin A &\implies \hat{s}_i \geq \frac{1}{\hat{m}} \\
&\implies \sum_{t=1}^{h} \mathbb{1}\{X_i = j\} \geq \frac{h}{\hat{m}}
\end{aligned}
$$

Let $Z_{ij} = \sum_{t=1}^{h/2} \mathbb{1}\{X_i = j\}$ and $t'_1 \ldots t'_{Z_{ij}} = t : X_{i,t} = j$

$$
\hat{\mu}_{i,j} = \frac{1}{Z_{ij}} \sum_{t'=1}^{Z_{ij}} r_{t'}
$$

$$
\begin{aligned}
\mathrm{P}\left(|\hat{\mu}_{i,j} - \mu_{ij}| > \epsilon\right) &= \sum_{z=1}^{\infty} \mathrm{P}\left(Z_{ij} = z\right) \mathrm{P}\left(\left|\frac{1}{Z_{ij}} \sum_{t'=1}^{Z_{ij}} r_{t'} - \mu_{ij}\right| > \epsilon | Z_{ij} = z\right) \\
&= \sum_{z=1}^{\infty} \mathrm{P}\left(Z_{ij} = z\right) \mathrm{P}\left(\left|\frac{1}{z} \sum_{t'=1}^{z} r_{t'} - \mu_{ij}\right| > \epsilon\right) \\
&\leq \mathrm{P}\left(\left|\frac{2\hat{m}}{h} \sum_{t'=1}^{h/2\hat{m}} r_{t'} - \mu_{ij}\right| > \epsilon\right) \sum_{z=1}^{\infty} \mathrm{P}\left(Z_{ij} = z\right) \\
&\leq 2 \exp - \frac{h\epsilon^2}{\hat{m}}
\end{aligned}
$$

Applying the union bound over all $2N$ actions.

$$\begin{aligned} &\mathrm{P}\left(\exists (i,j) : |\hat{\mu}_{i,j} - \mu_{i,j}| > \epsilon\right) \le 4N \exp -\frac{h\epsilon^2}{\hat{m}} \\ \implies &\mathrm{P}\left(\exists (i,j) : |\hat{\mu}_{i,j} - \mu_{i,j}| > \sqrt{\frac{\hat{m}}{h} \log \frac{4N}{\delta}}\right) \le \delta \end{aligned}$$

Now by Lemma 4,

$$h \ge 24m \log \frac{4N}{\delta} \implies \mathrm{P}\left(\hat{m} > 2m\right) \le \delta$$

Therefore if $h \ge 24m \log \frac{4N}{\delta}$, we have with probability at least $1 - 2\delta$ that

$$(\forall i, j) \qquad |\hat{\mu}_{i,j} - \mu_{i,j}| \le \sqrt{\frac{2m}{h} \log \frac{4N}{\delta}} \text{ and } \hat{m} \le 2m \tag{10}$$

Suppose $h < 24m \log \frac{4N}{\delta}$. Then

$$|\hat{\mu}_{i,j} - \mu_{i,j}| \le 1 \le \sqrt{\frac{2m}{h} \log \frac{4N}{\delta}} .$$

Therefore

$$|\hat{\mu}_{i,j} - \mu_{i,j}| \le \sqrt{\frac{24m}{h} \log \frac{4N}{\delta}} \qquad \forall h \text{ with probability at least } 1 - 2\delta$$

Therefore

$$\begin{aligned} R_s(h) &\le 2\delta + \sqrt{\frac{24m}{h} \log \frac{4N}{\delta}} \\ &\le \frac{8m}{h} + \sqrt{\frac{24m}{h} \log \left(\frac{Nh}{m}\right)} \end{aligned}$$

as required. □

3 Proof of Theorem 3

Theorem 5. *Define* $m = min\left\{1 \le i \le N : q_{i+1} \ge \frac{1}{i}\right\}$ *Then Algorithm 3 satisfies*

$$R(T) \in \mathcal{O}\left(T^{2/3} m^{1/3} log(KT)^{1/3}\right) .$$

- Use $24Nlog(4N/\delta)$ samples to estimate m
- If $\hat{m} > \lambda = \frac{N^{3/2}}{\sqrt{T}}$ then stop and do UCB with remaining rounds.
- Else use causal explore-exploit and let the exploration time $h = T^{2/3}\hat{m}^{1/3}log(TK)^{1/3}$

There will be 3 cases.

1. $m < \lambda/2 \implies \hat{m} < \lambda$ with prob $1 - \delta$

$$R_T \sim O(T^{2/3}m^{1/3}) + O(\sqrt{NT})\delta \implies \text{ want } \delta \leq \frac{T^{1/6}}{\sqrt{N}}$$

2. $m > 3\lambda/2 \implies \hat{m} > \lambda$ with prob $1 - \delta$

$$R_T \sim O(T^{2/3}N^{1/3})\delta + O(\sqrt{NT}) \implies \text{ want } \delta \leq \frac{N^{1/6}}{T^{1/6}}$$

3. $\lambda/2 < m < 3\lambda/2$, algorithm could end up doing UCB or Explore-Exploit.

$$\begin{aligned} R_T \sim & O(T^{2/3}m^{1/3}) + O(\sqrt{NT}) \\ = & O(\sqrt{NT}) \text{ as } m = O(\frac{N^{3/2}}{\sqrt{T}}) \end{aligned}$$

So if $\delta = \frac{1}{T^{1/6}\sqrt{N}}$ (or $\delta = \frac{1}{T^{1/3}}$) that would be enough concentration around m to choose the correct algorithm ... Is it enough to choose a reasonable total exploration time h? It seems that $\delta = \frac{1}{T^{1/3}}$ works.

Since we don't know $\boldsymbol{q}$, and thus m, we can't set the exploration time h in advance based on m as we did for the known $\boldsymbol{q}$ case. Instead we first use $24Nlog(4N/\delta)$ samples to estimate m. If $\hat{m} < \frac{N^{3/2}}{\sqrt{T}}$ we will continue with the causal algorithm and let the total exploration time $h = T^{2/3}\hat{m}^{1/3}log(2TK)^{1/3}$.

Since $m \leq N$, by Eq. (10), we have that with probability at least $1 - 2\delta$,

$$(\forall i, j) \qquad |\hat{\mu}_{i,j} - \mu_{i,j}| \leq \sqrt{\frac{\hat{m}}{h} \log \frac{4N}{\delta}} \text{ and } \hat{m} \leq 2m \tag{11}$$

So the regret in this case is bounded by,

$$\begin{aligned} R(T) &= 2\delta E\left[R(T) | \text{ not } Eq.\ (11)\right] + (1 - 2\delta)E[R(T)|Eq.\ (11)] \\ &\leq 2\delta T + \left(h + T\sqrt{\frac{\hat{m}}{h} \log \frac{4N}{\delta}}\right) \\ &\leq 2T^{2/3} + h + T\left(\sqrt{\frac{\hat{m}}{h} \log\left(4NT^{1/3}\right)}\right), \text{ letting } \delta = \frac{1}{T^{1/3}} \\ &\leq 2T^{2/3} + h + T\left(\sqrt{\frac{\hat{m}}{h} \log\left(2TK\right)}\right) \\ &\leq 2T^{2/3} + 2T^{2/3}(2m)^{1/3}log(2TK)^{1/3} \\ &\leq 6T^{2/3}m^{1/3}log(TK)^{1/3} \end{aligned}$$