Committing Bandits
(Supplementary Material)

1 Relevant policies and results

1.1 Allocation policies:

Uniform allocation (Unif): Plays all arms in the round robin fashion. Formally, for each time $t = 1, 2, \ldots$, set $I_t = t \mod K$.

Upper Confidence Bound (UCB) [1]: From time 1 to time K, pull each arm once. For time $t = K+1, K+2, \ldots$, pull the arm I_t such that

$$I_t \;=\; \arg\max_{1\le i\le K} \left(\hat{\theta}_{i,T_i(t-1)} + \sqrt{\frac{2\ln(t-1)}{T_i(t-1)}} \right),$$

where $\hat{\theta}_{i,T_i(t-1)}$ is the empirical average of rewards associated with arm i so far, i.e.,

$$\hat{\theta}_{i,T_i(t-1)} \;=\; \frac{1}{T_i(t-1)} \sum_{s=1}^{T_i(t-1)} X_{i,s}. \tag{1}$$

1.2 Recommendation policies:

Empirical Distribution of Plays (EDP): Recommend arm i with probability $T_i(n)/n$. That is,

$$\mathbb{P}(J_n = i) \;=\; \frac{T_i(n)}{n}.$$

Empirical Best Arm (EBA): Recommend the arm which achieves maximum empirical average of rewards so far, i.e.,

$$J_n \in \arg\max_{1\le i\le K} \hat{\theta}_{i,T_i(n)},$$

where $\hat{\theta}_{i,T_i(n)}$ is defined in (1).

Most Played Arm (MPA): Recommend the most played arm, i.e.,

$$J_n \in \arg\max_{1\le i\le K} T_i(n).$$

1.3 Known results

First, it is easy to see that $\mathbb{E}[R_n] \le \theta^* n$ for any allocation policy.

Result 1 (Distribution-dependent [5])**.** *For any allocation policy, and for any set of reward distributions such that their parameters θ_i are not all equal, there exists an ordering of $(\theta_1, \ldots, \theta_K)$ such that*

$$\mathbb{E}[R_n] \;\ge\; \left(\sum_{i\ne i^*} \frac{\Delta_i}{D(p_i \| p^*)} + o(1) \right) \ln n,$$

where $D(p_i \| p^*) = p_i \log \frac{p_i}{p^*} + p^* \log \frac{p^*}{p_i}$ is the Kullback-Leibler divergence between two Bernoulli reward distributions p_i (of arm i) and p^* (of the optimal arm), and $o(1) \to 0$ as $n \to \infty$.

Result 2 (Distribution-free [6])**.** *There exist positive constants c and N_0 such that for any allocation policy, there exists a set of Bernoulli reward distributions such that*

$$\mathbb{E}[R_n] \ \geq \ cK(\ln n - \ln K), \quad \forall n \geq N_0.$$

Result 3 (Distribution-dependent [3])**.** *For any pair of allocation and recommendation policies, if the allocation policy can achieve an upper-bound such that for all (Bernoulli) reward distributions $\theta_1, \ldots, \theta_K$, there exists a constant $C \geq 0$ with*

$$\mathbb{E}[R_n] \ \leq \ Cf(n),$$

then, for all sets of $K \geq 3$ Bernoulli reward distributions, with parameters θ_i that are all distinct and all different from 1, there exists an ordering $(\theta_1, \ldots, \theta_K)$ such that

$$\mathbb{E}[r_n] \ \geq \ \frac{\Delta}{2} e^{-Df(n)},$$

where D is a constant which can be calculated in closed form from C, and $\theta_1, \ldots, \theta_K$.

In particular, since $\mathbb{E}[R_n] \leq \theta^* n$ for any allocation policy, there exists a constant ξ depending only on $\theta_1, \ldots, \theta_K$ such that $\mathbb{E}[r_n] \geq (\Delta/2)e^{-\xi n}$.

Result 4 (Distribution-free [3])**.** *For any pair of allocation policy and any recommendation policy, there exists a set of Bernoulli reward distributions such that*

$$\mathbb{E}[r_n] \ \geq \ \frac{1}{20}\sqrt{\frac{K}{n}}.$$

Result 5 (Distribution-dependent [1])**.** *For the* **UCB** *allocation algorithm,*

$$\mathbb{E}_{UCB}[R_n] \ \leq \ \left(\sum_{i:\Delta_i > 0} \frac{8}{\Delta_i} + o(1) \right) \ln n,$$

where $o(1) \to 0$ as $n \to \infty$. Thus, by Result 3, for **UCB** *together with any recommendation policy, there exists a constant ρ such that $\mathbb{E}[r_n] \geq (\Delta/2)n^{-\rho}$.*

Result 6 (Distribution-dependent [3])**.** *Upper-bounds on simple regret:*

1. *For the pair* **[Unif, EBA]***, $\mathbb{E}[r_n] \leq \sum_{i \neq i^*} \Delta_i e^{-\Delta_i^2 \lfloor n/K \rfloor}$, for all $n \geq K$.*
2. *For the pair* **[UCB, MPA]***, $\mathbb{E}[r_n] \ \leq \ \frac{K^3}{(n-K)^2}$, for all n sufficiently large, such that $n \geq K + 4K \ln n / \Delta^2$ and $n \geq K(K+2)$.*

2 Theorem 1 and its proof

Theorem 1. *(1)* Distribution-dependent lower bound*: In Regime 1, for any algorithm, and any set of $K \geq 3$ Bernoulli reward distributions such that θ_i are all distinct and all different from 1, there exists an ordering $(\theta_1, \ldots, \theta_K)$ such that*

$$\mathbb{E}[\mathsf{Reg}] \ \geq \ \left(\max \left\{ \frac{(1-\gamma)\theta^*}{\xi}, \sum_{i \neq i^*} \frac{\Delta_i}{D(p_i \| p^*)} \right\} + o(1) \right) \frac{\ln T}{T},$$

where $o(1) \to 0$ as $T \to \infty$, and ξ is the constant discussed in Result 3.

(2) Distribution-free lower bound*: Also, for any algorithm in Regime 1, there exists a set of Bernoulli reward distributions such that*

$$\mathbb{E}[\mathsf{Reg}] \ \geq \ cK \left(1 - \frac{\ln K}{\ln T} \right) \frac{\ln T}{T},$$

where c is the constant in Result 2.

Proof. We first derive the distribution-dependent lower-bound. Combining two lower bounds in Results 1 and 3 yields that

$$\begin{aligned}\mathbb{E}[\mathsf{Reg}] &\geq \frac{\gamma c_1 \ln N}{T} + \frac{T-N}{T} c_2 e^{-c_3 N} + \frac{(1-\gamma)N\theta^*}{T} \\ &\geq \frac{1}{T}\left((T-N)c_2 e^{-c_3 N} + (1-\gamma)N\theta^*\right),\end{aligned}$$

where $c_1 = \sum_{i \neq i^*} \Delta_i / D(p_i \| p^*)$; $c_2 = \Delta/2$; and $c_3 = \xi$. Now, let $F_0(N) := (1-\gamma)N\theta^* + (T-N)c_2 e^{-c_3 N}$. We have that $F_0(N)$ is convex for $N \in [0, T]$, and

$$\frac{\partial F_0}{\partial N} = (1-\gamma)\theta^* - c_2 e^{-c_3 N}(1 + c_3(T-N)).$$

Thus defining N^* by $\partial F_0(N^*)/\partial N = 0$, we have:

$$\frac{(1-\gamma)\theta^*}{c_2 c_3} e^{c_3 N^*} + N^* - \frac{1}{c_3} = T. \tag{2}$$

With T going to infinity, N^* also goes to infinity, and hence, the first term in (2) dominates the second term if T is large enough. Therefore, for T large enough,

$$\begin{aligned}T = \frac{(1-\gamma)\theta^*}{c_2 c_3} e^{c_3 N^*} + N^* - \frac{1}{c_3} &\leq 2\frac{(1-\gamma)\theta^*}{c_2 c_3} e^{c_3 N^*} \\ \text{i.e.} \qquad N^* &\geq \frac{1}{c_3}\left(\ln T - \ln\left(\frac{c_2 c_3}{2(1-\gamma)\theta^*}\right)\right).\end{aligned}$$

Substituting (2) into F_0, we obtain:

$$\begin{aligned}F_0(N) &\geq F_0(N^*) = (1-\gamma)\theta^*\left(N^* + \frac{1}{c_3} - \frac{c_2}{c_3(1-\gamma)\theta^*} e^{-c_3 N^*}\right) \\ &\geq (1-\gamma)\theta^*\left(\frac{\ln T}{c_3} - \frac{1}{c_3}\ln\left(\frac{c_2 c_3}{2(1-\gamma)\theta^*}\right) + \frac{1}{c_3} - \frac{2}{c_3^2 T}\right).\end{aligned}$$

Therefore,

$$\mathbb{E}[\mathsf{Reg}] \geq \left(\frac{(1-\gamma)\theta^*}{\xi} + o(1)\right)\frac{\ln T}{T}, \tag{3}$$

where $o(1) \to 0$ as $T \to \infty$.

Alternatively, we note that

$$\begin{aligned}\mathbb{E}[\mathsf{Reg}] &= \gamma\frac{\mathbb{E}[R_N]}{T} + \frac{T-N}{T}\mathbb{E}[r_N] + (1-\gamma)\frac{N}{T}\theta^* \\ &\geq \frac{1}{T}\left(\mathbb{E}[R_N] + (T-N)\mathbb{E}[r_N]\right),\end{aligned}$$

since $\mathbb{E}[R_N] \leq \theta^* N$. But the right hand side is nothing but the regret of a particular strategy for the usual multi-armed bandit problem in T slots, and hence, it is further lower-bounded by Result 1. Thus,

$$\mathbb{E}[\mathsf{Reg}] \geq \left(\sum_{i \neq i^*}\frac{\Delta_i}{D(p_i \| p^*)} + o(1)\right)\frac{\ln T}{T}. \tag{4}$$

Combining (3) and (4) yields the first bound.

Now, the distribution-free lower-bound can be obtained by noticing the following:

$$\begin{aligned}\mathbb{E}[\mathsf{Reg}] &= \gamma\frac{\mathbb{E}[R_N]}{T} + \frac{T-N}{T}\mathbb{E}[r_N] + (1-\gamma)\frac{N}{T}\theta^* \\ &\geq \frac{1}{T}\left(\mathbb{E}[R_N] + (T-N)\mathbb{E}[r_N]\right),\end{aligned}$$

since $\mathbb{E}[R_N] \leq \theta^* N$. As we claimed before, the right hand side is nothing but the regret of a particular strategy for the usual multi-armed bandit problem in T slots, and hence, it is further lower-bounded by Result 2. Thus, $\mathbb{E}[\mathsf{Reg}] \geq cK/T(\ln T - \ln K)$, for all $T \geq N_0$, where c and N_0 are defined in Result 2. The result then follows. $\square$

3 Theorem 2 and its proof

Theorem 2. *For the* **Non-adaptive Unif-EBA** *algorithm,*

$$\mathbb{E}[\mathsf{Reg}] \;\leq\; \frac{K}{\Delta^2}\left((1-\gamma)\theta^* + \frac{\gamma}{K}\sum_{i\neq i^*}\Delta_i + \frac{2\Delta^2}{\ln T}\right)\frac{\ln T}{T}.$$

Proof. The proof follows immediately from the known upper bound of the pair **[Unif, EBA]** (see [3]). Since the algorithm chooses uniformly each arm $\lceil \ln T/\Delta^2 \rceil \leq 1 + \ln T/\Delta^2$ times, we have that

$$N \;\leq\; K\left(\frac{\ln T}{\Delta^2}+1\right), \quad \mathbb{E}[R_N] \;\leq\; \sum_{i\neq i^*}\Delta_i\left(\frac{\ln T}{\Delta^2}+1\right),$$

$$\mathbb{E}[r_N] \;\leq\; \sum_{i\neq i^*}\Delta_i e^{-\Delta_i^2(\ln T/\Delta^2)} \;\leq\; \sum_{i\neq i^*}\Delta_i\frac{1}{T}.$$

Therefore,

$$\begin{aligned}
\mathbb{E}[\mathsf{Reg}] &= \frac{(1-\gamma)\theta^*}{T}N + \frac{\gamma}{T}\mathbb{E}[R_N] + \frac{T-N}{T}\mathbb{E}[r_N] \\
&\leq \frac{K}{\Delta^2}\left((1-\gamma)\theta^* + \frac{\gamma}{K}\sum_{i\neq i^*}\Delta_i + \frac{\Delta^2}{\ln T}\left((1-\gamma)\theta^* + \sum_{i\neq i^*}\frac{\Delta_i}{K}(\gamma+1)\right)\right)\frac{\ln T}{T} \\
&\leq \frac{K}{\Delta^2}\left((1-\gamma)\theta^* + \frac{\gamma}{K}\sum_{i\neq i^*}\Delta_i + \frac{2\Delta^2}{\ln T}\right)\frac{\ln T}{T},
\end{aligned}$$

where the last inequality is due to the facts that $0 < \theta^* \leq 1$ and $0 < \Delta_i \leq 1$. ☐

4 Theorem 3 and its proof

Theorem 3. *(1)* Distribution-dependent lower bound*: In Regime 2, for any algorithm, and any set of $K \geq 3$ Bernoulli reward distribution such that θ_i are all distinct and all different from 1, there exists an ordering $(\theta_1, \ldots, \theta_K)$ such that*

$$\mathbb{E}[\mathsf{Reg}] \;\geq\; \left(\sum_{i\neq i^*}\frac{\Delta_i}{D(p_i\|p^*)} + o(1)\right)\frac{\ln T}{T},$$

where $o(1) \to 0$ as $T \to \infty$.

(2) Distribution-free lower bound*: Also, for any algorithm in Regime 2, there exists a set of Bernoulli reward distributions such that*

$$\mathbb{E}[\mathsf{Reg}] \;\geq\; cK\left(1 - \frac{\ln K}{\ln T}\right)\frac{\ln T}{T},$$

where c is the constant from Result 2.

Proof. Given a fixed N, using the same technique as in the proof of Theorem 1, we note that:

$$\begin{aligned}
\mathbb{E}[\mathsf{Reg}|N] &= \gamma\frac{\mathbb{E}[R_N|N]}{T} + \frac{T-N}{T}\mathbb{E}[r_N|N] + (1-\gamma)\frac{N}{T}\theta^* \\
&\geq \frac{1}{T}\left(\mathbb{E}[R_N|N] + (T-N)\mathbb{E}[r_N|N]\right),
\end{aligned}$$

since $\mathbb{E}[R_N|N] \leq \theta^* N$. However, any algorithm in Regime 2 is just a particular algorithm for the usual stochastic multi-armed bandit problem in T slots, and the right hand side is nothing but its regret. Therefore, applying the distribution-dependent lower bound in Result 1 and the distribution-free lower bound in Result 2, and taking the expectation over N yield the results. ☐

5 Theorem 4 and its proof

Theorem 4. *For the* **SEC1** *algorithm,*

$$\mathbb{E}[\mathsf{Reg}] \;\le\; \frac{K}{\Delta^2}\left((1-\gamma)\theta^* + \frac{\gamma}{K}\sum_{i\neq i^*}\Delta_i + b\right)\frac{\ln T}{T},$$

where

$$b = \left(2 + \frac{\Delta^2(K+2)}{(1-e^{-\Delta^2/2})^2}\right)\frac{1}{\ln T} \;\to\; 0 \quad \textit{as} \quad T \;\to\; \infty.$$

Proof. Let $P_k^*(n)$ denote the probability that the optimal arm i^* is kept until the end of n-th round, and $P_r^*(n)$ denote the probability that i^* is rejected within the first n rounds. By definition, $P_k^*(n) = 1 - P_r^*(n)$.

For $n \le \alpha \ln T$, using the "maximal" version of Chernoff-Hoeffding bound [4], we have that

$$P_r^*(n) = \mathbb{P}\left(\max_{1\le j\le n}\left|S_j^{i^*} - j\theta^*\right| > \epsilon_1 \ln T\right) \le 2e^{-2\epsilon_1^2(\ln T)^2/n} \le 2e^{-2\epsilon_1^2(\ln T)/\alpha} = 2T^{-2}, \qquad (5)$$

since $\alpha = 1/\Delta^2$, $\epsilon_1 = 1/\Delta$.

For $\alpha \ln T < n \le T$, using the union bound, Chernoff-Hoeffding bound [4] and (5), we have that:

$$\begin{aligned}
P_r^*(n) &\le \mathbb{P}(\text{“}i^* \text{ is rejected before } \alpha\ln T\text{”}) + \sum_{j=\alpha\ln T}^{n}\mathbb{P}(\text{“arm } i^* \text{ is rejected at time } j\text{”}) \\
&= P_r^*(\alpha\ln T) + \sum_{j=\alpha\ln T}^{n}\mathbb{P}(|S_j^{i^*} - j\theta^*| > \epsilon_2 j) \\
&\le 2T^{-2} + \sum_{j=\alpha\ln T}^{n} 2e^{-2\epsilon_2^2 j} \;=\; 2T^{-2} + \frac{2e^{-2\epsilon_2^2\alpha\ln T}}{1-e^{-2\epsilon_2^2}}(1-e^{-2\epsilon_2^2(n-\alpha\ln T)}) \\
&\le 2T^{-2\epsilon_1^2/\alpha} + \frac{2T^{-2\epsilon_2^2\alpha}}{1-e^{-2\epsilon_2^2}} \;=\; 2T^{-2} + \frac{2T^{-1/2}}{1-e^{-\Delta^2/2}}, \qquad (6)
\end{aligned}$$

since $\epsilon_2 = \Delta/2$.

Now, consider any arm i that is not optimal, and let $P_k^i(n)$ denote the probability that arm i is kept until the end of n-th round. Let $h(n)$ denote the rejecting threshold at time n, i.e., $h(n) = \epsilon_1 \ln T$ for $n \le \alpha \ln T$, and $h(n) = \epsilon_2 n$ for $n > \alpha \ln T$. Then for $n > \alpha \ln T$,

$$\begin{aligned}
P_k^i(n) &= \mathbb{P}\left(\left|S_j^i - j\theta^*\right| \le h(j),\ 1 \le j \le n\right) \\
&\le \mathbb{P}\left(\left|S_n^i - n\theta^*\right| \le \epsilon_2 n\right) \\
&= \mathbb{P}\left(-\epsilon_2 n \le S_n^i - n\Delta_i - n\theta_i \le \epsilon_2 n\right) \\
&\le \mathbb{P}\left(S_n^i - n\theta_i \ge n(\Delta_i - \epsilon_2)\right) \;\le\; e^{-n\Delta^2/2}.
\end{aligned}$$

Therefore, for $n > \alpha \ln T$, the probability that any suboptimal arm is kept until the end of n-th round at most $(K-1)e^{-n\Delta^2/2}$, and hence, the probability that all suboptimal arms are rejected within first n rounds is at least $1-(K-1)e^{-n\Delta^2/2}$.

Let N^s denote the stopping time of the algorithm. Then, for $n > \alpha \ln T$,

$$\mathbb{P}\left(N^s \le n\right) \ge \mathbb{P}\left(\text{all suboptimal arms are rejected within first } n \text{ rounds}\right) \ge 1-(K-1)e^{-n\Delta^2/2}.$$

$$\Rightarrow \quad \mathbb{P}\left(N^s > n\right) \;\le\; (K-1)e^{-n\Delta^2/2} \quad \text{for } n > \alpha\ln T.$$

And hence,

$$
\begin{aligned}
\mathbb{E}\left[N^s\right] \;=\; \sum_{1\le n\le T} \mathbb{P}\left(N^s > n\right) \;&\le\; \alpha \ln T + \sum_{\alpha \ln T < n \le T} (K-1)e^{-n\Delta^2/2} \\
&=\; \alpha \ln T + \frac{(K-1)T^{-\alpha\Delta^2/2}}{1-e^{-\Delta^2/2}}(1-e^{-T\Delta^2/2}) \\
&\le\; \frac{\ln T}{\Delta^2} + \frac{(K-1)}{1-e^{-\Delta^2/2}} T^{-1/2} \\
&\le\; \frac{\ln T}{\Delta^2}\left(1 + \frac{\Delta^2(K-1)}{(1-e^{-\Delta^2/2})\ln T}\right). \qquad (7)
\end{aligned}
$$

Thus, the cumulative regret bound is:

$$
\mathbb{E}\left[R_{N_s}\right] \;\le\; \gamma\left(\sum_{i\neq i^*}\Delta_i\right)\frac{\ln T}{\Delta^2}\left(1+\frac{\Delta^2(K-1)}{(1-e^{-\Delta^2/2})\ln T}\right). \qquad (8)
$$

Now, let us consider the simple regret:

$$
\begin{aligned}
\mathbb{E}[r_{N_s}] \;&=\; \sum_{i\neq i^*}\Delta_i\,\mathbb{P}(\text{“arm } i \text{ is kept until the stopping time”}) \\
&=\; \sum_{i\neq i^*}\sum_{n=1}^{T}\Delta_i\mathbb{P}(\text{“arm } i \text{ is kept until } n\text{”},\; N_s = n) \\
&\le\; \sum_{i\neq i^*}\sum_{n=1}^{T}\Delta_i\mathbb{P}(\text{“arm } i \text{ is kept until } n\text{”},\; \text{“arm } i^* \text{ is rejected before } n\text{”}) \\
&=\; \sum_{i\neq i^*}\sum_{n=1}^{T}\Delta_i\mathbb{P}(\text{“arm } i \text{ is kept until } n\text{”}) \times \mathbb{P}(\text{“arm } i^* \text{ is rejected before } n\text{”}). \\
&=\; \sum_{i\neq i^*}\Delta_i\sum_{n=1}^{T}P_k^i(n)P_r^*(n) \\
&=\; \sum_{i\neq i^*}\Delta_i\left(\sum_{1\le n\le\alpha\ln T}P_k^i(n)P_r^*(n) + \sum_{\alpha\ln T<n\le T}P_k^i(n)P_r^*(n)\right),
\end{aligned}
$$

where the fourth equality is because **SEC1** makes decision on each arm independently. Then, applying (6) and (5) yields that

$$
\begin{aligned}
\mathbb{E}[r] \;&\le\; \sum_{i\neq i^*}\Delta_i\left(\sum_{1\le n\le\alpha\ln T}2T^{-2} + \sum_{\alpha\ln T<n\le T}e^{-n\Delta^2/2}\left(T^{-2}+\frac{2T^{-1/2}}{1-e^{-\Delta^2/2}}\right)\right) \\
&\le\; \sum_{i\neq i^*}\Delta_i\left(2\alpha(\ln T)T^{-2} + \frac{T^{-1/2}}{1-e^{-\Delta^2/2}}\left(T^{-2}+\frac{2T^{-1/2}}{1-e^{-\Delta^2/2}}\right)\right) \\
&\le\; \left(\sum_{i\neq i^*}\Delta_i\right)\left(\frac{2}{\Delta^2}T^{-2+1/e} + \frac{T^{-5/2}}{1-e^{-\Delta^2/2}} + \frac{2T^{-1}}{(1-e^{-\Delta^2/2})^2}\right) \\
&\le\; \left(\sum_{i\neq i^*}\Delta_i\right)\frac{1}{T(1-e^{-\Delta^2/2})^2}\left(3+\frac{2(1-e^{-\Delta^2/2})^2}{\Delta^2}\right), \qquad (9)
\end{aligned}
$$

where the third inequality is due to the fact that $\ln x \le x^{1/e}$ for all $x > 0$. Combining (7)-(9) yields that:

$$
\begin{aligned}
\mathbb{E}[\mathsf{Reg}] \;\le\;& \frac{K}{\Delta^2}\frac{\ln T}{T}\Bigg((1-\gamma)\theta^* + \frac{\gamma}{K}\sum_{i\neq i^*}\Delta_i \\
&+\frac{\Delta^2}{(1-e^{-\Delta^2/2})^2\ln T}\left[(1-\gamma)\theta^*(K-1)\left(1-e^{-\Delta^2/2}\right)\right] \\
&+\frac{\Delta^2}{(1-e^{-\Delta^2/2})^2\ln T}\left[\frac{\gamma(K-1)}{K}\left(\sum_{i\neq i^*}\Delta_i\right)\left(1-e^{-\Delta^2/2}\right)\right] \\
&+\frac{\Delta^2}{(1-e^{-\Delta^2/2})^2 K\ln T}\left(\sum_{i\neq i^*}\Delta_i\right)\left[3+\frac{2}{\Delta^2}\left(1-e^{-\Delta^2/2}\right)^2\right]\Bigg) \\
\le\;& \frac{K}{\Delta^2}\frac{\ln T}{T}\Bigg((1-\gamma)\theta^* + \frac{\gamma}{K}\sum_{i\neq i^*}\Delta_i \\
&+\frac{\Delta^2}{(1-e^{-\Delta^2/2})^2\ln T}\left[(K-1)+3+\frac{2}{\Delta^2}\left(1-e^{-\Delta^2/2}\right)^2\right]\Bigg) \\
=\;& \frac{K}{\Delta^2}\frac{\ln T}{T}\left((1-\gamma)\theta^* + \frac{\gamma}{K}\sum_{i\neq i^*}\Delta_i + \frac{1}{\ln T}\left[2+\frac{\Delta^2(K+2)}{(1-e^{-\Delta^2/2})^2}\right]\right),
\end{aligned}
$$

where the second inequality is due to the facts that $0 < \theta^* \le 1, 0 < \Delta_i \le 1$, and $e^{-\Delta^2/2} \ge 0$. $\square$

6 Theorem 5 and its proof

Theorem 5. *For the* **SC-UCB** *algorithm,*

$$
\mathbb{E}[\mathsf{Reg}] \;\le\; \sum_{i\neq i^*}\left(\frac{\gamma\Delta_i + (1-\gamma)\theta^*}{\Delta_i^2}\right)\frac{\ln(T\Delta_i^2)}{T}\left(32+\frac{\Delta_i^2+96}{\ln(T\Delta_i^2)}\right).
$$

Proof. This proof is based on the proof of Theorem 3.1 in [2].

First, let N_e^i denote the number of time slots that arm i is chosen during experimentation phase (i.e., before commitment), and N_c^i denote the number of time slots that i is chosen during commitment phase (that is, $N_c^j = 0$ if the algorithm does not commit to arm j). Also, let $N^i = N_e^i + N_c^i$ be the total number of time slots that the algorithm spent on arm i.

We then observe that in Regime 2, in any time slot *before* commitment, if the algorithm chooses a suboptimal arm i, then it suffers an expected loss $\gamma\Delta_i + (1-\gamma)\theta^*$; otherwise, it suffers an expected loss $(1-\gamma)\theta^*$ if it chooses i^*. Furthermore, in any time slot *after* commitment, if the algorithm commits to a suboptimal arm i, then it suffers an expected loss Δ_i; otherwise, it does not suffer any loss if committing to i^*.

Now, let us define $\lambda = \sqrt{e/T}$, let A be the set of arms i for which $\Delta_i > \lambda$, i.e., $A = \{i \in [K] : \Delta_i > \lambda\}$. For any arm $i \in A$, its contributed regret is

$$
\frac{(\gamma\Delta_i + (1-\gamma)\theta^*)\,N_e^i + \Delta_i N_c^i}{T} \le (\gamma\Delta_i + (1-\gamma)\theta^*)\,\frac{N_e^i + N_c^i}{T} = (\gamma\Delta_i + (1-\gamma)\theta^*)\,\frac{N^i}{T}.
$$

Following the steps in the proof of Theorem 3.1 in [2], one can show that the expected number of time slots that **SC-UCB** spent on arm i for $i \in A$ is bounded by:

$$
\mathbb{E}[N^i] \;\le\; \left(1+\frac{32\ln(T\Delta_i^2)}{\Delta_i^2}+\frac{96}{\Delta_i^2}\right), \qquad i \in A.
$$

Thus, the expected regret contributed by an arm $i \in A$ is bounded by

$$\frac{\gamma\Delta_i + (1-\gamma)\theta^*}{T}\left(1 + \frac{32\ln(T\Delta_i^2)}{\Delta_i^2} + \frac{96}{\Delta_i^2}\right).$$

Next, for $i \notin A$, we have that $N_e^i \leq 1 + 2\ln(T\lambda^2)/\lambda^2$ under **SC-UCB**, since **SC-UCB** will stop when $m = \lfloor \log_2(T/e)/2 \rfloor$. Moreover, N_c^i is trivially bounded by T. Thus, the contributed regret of an arm $i \notin A$ is bounded by

$$\frac{\gamma\Delta_i + (1-\gamma)\theta^*}{T}\left(\frac{2\ln(T\lambda^2)}{\lambda^2} + 1\right) + \Delta_i.$$

Therefore, the total regret bound would be

$$\begin{aligned}
\mathbb{E}[\mathsf{Reg}] \;&\leq\; \sum_{i:\Delta_i>\lambda} \frac{\gamma\Delta_i + (1-\gamma)\theta^*}{T}\left(1 + \frac{32\ln(T\Delta_i^2)}{\Delta_i^2} + \frac{96}{\Delta_i^2}\right) \\
&\quad + \sum_{i:\Delta_i\leq\lambda} \Delta_i + \frac{\gamma\Delta_i + (1-\gamma)\theta^*}{T}\left(1 + \frac{2\ln(T\lambda^2)}{\lambda^2}\right) \\
&\leq\; \sum_{i:\Delta_i>\lambda} \frac{\gamma\Delta_i + (1-\gamma)\theta^*}{T}\frac{\ln(T\Delta_i^2)}{\Delta_i^2}\left(32 + \frac{\Delta_i^2}{\ln(T\Delta_i^2)} + \frac{96}{\ln(T\Delta_i^2)}\right) \\
&\quad + \sum_{i:\Delta_i\leq\lambda} \frac{\gamma\Delta_i + (1-\gamma)\theta^*}{T}\frac{\ln(T\Delta_i^2)}{\Delta_i^2}\left(2 + \frac{\Delta_i^2}{\ln(T\Delta_i^2)} + \frac{T\Delta_i}{\gamma\Delta_i + (1-\gamma)\theta^*}\frac{\Delta_i^2}{\ln(T\Delta_i^2)}\right).
\end{aligned}$$

Note that for $\Delta_i \leq \lambda = \sqrt{e/T}$, $T\Delta_i^2 \leq e$. Moreover, $\Delta_i \leq \gamma\Delta_i + (1-\gamma)\theta^*$. Thus,

$$\begin{aligned}
\mathbb{E}[\mathsf{Reg}] \;&\leq\; \sum_{i:\Delta_i>\lambda} \frac{\gamma\Delta_i + (1-\gamma)\theta^*}{T}\frac{\ln(T\Delta_i^2)}{\Delta_i^2}\left(32 + \frac{\Delta_i^2}{\ln(T\Delta_i^2)} + \frac{96}{\ln(T\Delta_i^2)}\right) \\
&\quad + \sum_{i:\Delta_i\leq\lambda} \frac{\gamma\Delta_i + (1-\gamma)\theta^*}{T}\frac{\ln(T\Delta_i^2)}{\Delta_i^2}\left(2 + \frac{\Delta_i^2}{\ln(T\Delta_i^2)} + \frac{e}{\ln(T\Delta_i^2)}\right) \\
&\leq\; \sum_{i\neq i^*}\left(\frac{\gamma\Delta_i + (1-\gamma)\theta^*}{\Delta_i^2}\right)\frac{\ln(T\Delta_i^2)}{T}\left(32 + \frac{\Delta_i^2 + 96}{\ln(T\Delta_i^2)}\right).
\end{aligned}$$

□

7 Theorem 6 and its proof

Theorem 6. *The cumulative regret of* **UCB-poly**(δ) *is upper-bounded by*

$$\mathbb{E}[R_n] \;\leq\; \left(\sum_{i:\Delta_i>0}\frac{8}{\Delta_i} + o(1)\right)n^\delta,$$

where $o(1) \to 0$ as $n \to \infty$. Moreover, the simple regret for the pair **[UCB-poly**(δ)**, EBA]** *is upper-bounded by*

$$\mathbb{E}[r_n] \;\leq\; \left(2\sum_{i\neq i^*}\Delta_i\right)e^{-\chi n^\delta},$$

where $\chi = \min_i \frac{\sigma}{2}\Delta_i^2$.

Proof. Following the steps in the proof of Theorem 1 in [1], we can easily show the following cumulative regret bound for **UCB-poly**(δ):

$$\begin{aligned}
\mathbb{E}_{\text{UCB-poly}}[T_i(n)] \;&\leq\; \left\lceil\frac{8n^\delta}{\Delta_i^2}\right\rceil + \sum_{t=1}^{n}\sum_{s=1}^{t}\sum_{s_i=1}^{t} 2e^{-4t^\delta} \;\leq\; \frac{8n^\delta}{\Delta_i^2} + 1 + 2\sum_{t=1}^{n} t^2 e^{-4t^\delta} \\
&\leq\; \frac{8n^\delta}{\Delta_i^2} + 1 + 2A^\infty(\delta),
\end{aligned}$$

where $A^{\infty}(\delta) = \sum_{t=1}^{\infty} t^2 e^{-4t^{\delta}}$. Since $A^{\infty}(\delta)$ is finite for any fixed $\delta > 0$, the result is obtained.

In order to prove the simple regret bound for the pair **[UCB-poly(δ), EBA]**, we need the following lemma.

Lemma 1. *There exists a positive constant σ such that under the* **UCB-poly(δ)** *policy, for any arm i and any $n > K$,*

$$T_i(n) \;\geq\; \sigma n^{\delta}.$$

Proof. We first note that for any i, $T_i(n) \geq 1$ for any $n > K$, and $T_i(n)$ is non-decreasing in n. Therefore, if such constant σ does not exist, it has to be the case in which there exists an arm j such that $T_j(n)/n^{\delta} \to 0$ as $n \to \infty$, or $T_j(n) \ll n^{\delta}$. It is then elementary to prove that this cannot happen under the **UCB-poly(δ)** policy. □

Now we prove the simple regret bound the pair **[UCB-poly(δ), EBA]**. Note that

$$\begin{aligned}
\mathbb{E}[r_n] &= \mathbb{E}\left[\theta^* - \theta_{J_n}\right] \;=\; \mathbb{E}\left[\sum_{i \neq i^*} \Delta_i \{J_n = i\}\right] \\
&= \sum_{i \neq i^*} \Delta_i \mathbb{P}\left(J_n = i\right) \;\leq\; \sum_{i \neq i^*} \Delta_i \mathbb{P}\left(\hat{\theta}_{i,T_i(t)} \geq \hat{\theta}^*_{T^*(t)}\right).
\end{aligned}$$

If $\hat{\theta}_{i,T_i(t)} < \theta_i + \frac{\Delta_i}{2}$ and $\hat{\theta}^*_{T^*(t)} > \theta^* - \frac{\Delta_i}{2}$, then $\hat{\theta}_{i,T_i(t)} < \hat{\theta}^*_{T^*i(t)}$. Thus,

$$\begin{aligned}
\mathbb{P}\left(\hat{\theta}_{i,T_i(t)} \geq \hat{\theta}^*_{T^*(t)} \,\middle|\, T_i(t), T^*(t)\right) &\leq \mathbb{P}\left(\hat{\theta}_{i,T_i(t)} \geq \theta_i + \frac{\Delta_i}{2}\right) + \mathbb{P}\left(\hat{\theta}^*_{T^*(t)} \leq \theta^* - \frac{\Delta_i}{2}\right) \\
&\leq \exp\left(-\frac{\Delta_i^2 T_i(t)}{2}\right) + \exp\left(-\frac{\Delta_i^2 T^*(t)}{2}\right) \\
&\leq 2\exp\left(-\frac{\sigma \Delta_i^2}{2} n^{\delta}\right),
\end{aligned}$$

where the second inequality is due to the Chernoff-Hoeffding bound [4], and the third inequality is due to Lemma 1. Taking the expectation over $(T_i(t), T^*(t))$ yields that

$$\mathbb{P}\left(\hat{\theta}_{i,T_i(t)} \geq \hat{\theta}^*_{T^*(t)}\right) \;\leq\; 2\exp\left(-\frac{\sigma \Delta_i^2}{2} n^{\delta}\right).$$

Therefore,

$$\mathbb{E}[r_n] \;\leq\; 2\sum_{i \neq i^*} \Delta_i \exp\left(-\frac{\sigma \Delta_i^2}{2} n^{\delta}\right) \;\leq\; \left(2\sum_{i \neq i^*} \Delta_i\right) e^{-\chi n^{\delta}},$$

where $\chi = \min_i \dfrac{\sigma}{2} \Delta_i^2$. □

8 Theorem 7 and its proof

Theorem 7. *Suppose $T \equiv T(N)$ is a continuous function of N such that $\lim_{N \to \infty} \dfrac{\ln(\ln(T(N) - N))}{\ln N}$ exists. Consider the function*

$$F_{\delta}(N) \;:=\; C_1 \frac{N^{\delta}}{T(N)} + \frac{T(N) - N}{T(N)} C_2 e^{-C_3 N^{\delta}}, \quad 0 \leq \delta \leq 1,$$

where C_1, C_2, and C_3 are positive constants, and let

$$\delta^* := \lim_{N \to \infty} \frac{\ln(\ln(T(N) - N))}{\ln N}, \textit{ projected on } [0,1].$$

Then, for any $\delta \in [0,1]$, $\limsup_{N \to \infty} \dfrac{F_{\delta^*}(N)}{F_{\delta}(N)} \leq 1.$

The above theorem show that in the limit, as T and N increase to infinity, the "optimal" value of δ can be chosen as $\lim_{N\to\infty} \ln(\ln(T(N)-N))/\ln N$ if that limit exists. For examples:

- If $T(N) \doteq \Omega(e^N)$, then we should choose $\delta = 1$. The corresponding scheme is **[Unif, EBA]**, i.e., to explore uniformly during the experimentation phase, and then commit to the empirical best arm.
- If $T(N) \doteq \Theta(e^{N^\alpha})$ $(0 < \alpha < 1)$, then we should choose $\delta = \alpha$. The corresponding scheme is **[UCB-poly(α), EBA]**, i.e., to use the UCB-poly(α) algorithm during the experimentation phase, and then commit to the empirical best arm.
- If $T(N) \doteq \mathcal{O}(e^{\ln N})$, then we should choose $\delta = 0$. The corresponding scheme is **[UCB, MPA]**, i.e., to use the standard UCB algorithm during the experimentation phase, and then commit to the most played arm. Note that as $\delta = 0$, we cannot get the cumulative regret bound of $N^\delta = N^0 =$ constant, since $\ln N$ is a lower bound (and the standard UCB algorithm achieves that).

Proof. For a fixed N, let us define an auxiliary variable $y = N^\delta$ for $\delta \in [0,1]$ and define

$$F(y) \;:=\; C_1 \frac{y}{T(N)} + \frac{T(N)-N}{T(N)} C_2 e^{-C_3 y}, \qquad 1 \le y \le N.$$

Note that $F(y)$ is convex for $y \in [1,N]$. Therefore, it achieves a unique minimum at

$$y_N^* \;=\; \frac{1}{C_3} \ln\left(\frac{C_2 C_3 (T(N)-N)}{C_1}\right), \text{ projected on } [1,N],$$

or

$$\delta_N^* \;=\; \frac{1}{\ln N}\left(\ln\left(\ln(T(N)-N) + \ln\left(\frac{C_2 C_3}{C_1}\right)\right) - \ln C_3\right), \text{ projected on } [0,1]. \tag{10}$$

In other words, for any $\delta \in [0,1]$, we have that

$$F_{\delta_N^*}(N) \;\le\; F_\delta(N) \quad \text{or} \quad \frac{F_{\delta_N^*}(N)}{F_\delta(N)} \;\le\; 1 \quad \text{for all } N.$$

Also, taking the limit of (10) as N goes to infinity yields that

$$\lim_{N\to\infty} \delta_N^* = \lim_{N\to\infty} \frac{\ln(\ln(T(N)-N))}{\ln N} = \delta^*.$$

Now, suppose there exists some $\delta \in [0,1]$ such that

$$\limsup_{N\to\infty} \frac{F_{\delta^*}(N)}{F_\delta(N)} = 1 + \epsilon > 1, \quad \text{for some } \epsilon > 0.$$

Then there exists an increasing subsequence $\{n_0, n_1, n_2, \ldots\}$ such that

$$F_{\delta^*}(n_k) = (1+\epsilon) F_\delta(n_k) \ge (1+\epsilon) F_{\delta^*_{n_k}}(n_k) > F_{\delta^*_{n_k}}(n_k), \quad \text{for all } k.$$

This leads to a contradiction of the fact that $\lim_{N\to\infty} \delta_N^* = \delta^*$. Therefore,

$$\limsup_{N\to\infty} \frac{F_{\delta^*}(N)}{F_\delta(N)} \;\le\; 1, \quad \text{for any } \delta \in [0,1].$$

□

9 Additional simulation results

In this section, we present some additional numerical results on the performance of **Non-adaptive Unif-EBA**, **SEC1**, **SEC2**, and **SC-UCB** algorithms with different sets of parameters.

We first recall Figure 1 which shows the regrets of the above algorithms for various values of T (in

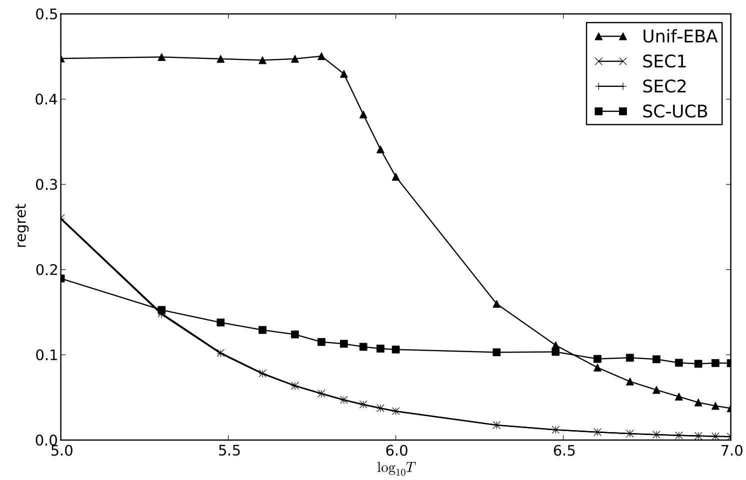

Figure 1: Numerical performances where $K = 20, \gamma = 0.75$, and $\Delta = 0.02$

logarithmic scale) with parameters $K = 20, \gamma = 0.75$, and $\Delta = 0.02$. We can see that the performances of **SEC1** and **SEC2** are nearly identical, which suggests that the requirement of knowing θ^* in **SEC1** can be relaxed. Moreover, **SEC1** (or equivalently, **SEC2**) performs much better than **Non-adaptive Unif-EBA** due to its adaptive nature. Particularly, the performance of **Non-adaptive Unif-EBA** is quite poor when the experimentation deadline is roughly equal to T, since the algorithm does not commit before the experimentation deadline. Finally, **SC-UCB** performs not as well as the others when T is large, but this algorithm does not require us to know Δ, and thus suffers a performance loss due to the additional effort required to estimate Δ.

Next, we keep $K = 20, \gamma = 0.75$, and decrease the value of Δ. Figures 2 and 3 show the performances of these algorithms for $\Delta = 0.01$ and $\Delta = 0.005$, respectively. The decrease of Δ affects all of the algorithms, but it seems to have more effect on **SEC1** and **SEC2** than **SC-UCB**, and particularly has a huge effect on the performance of **Non-adaptive Unif-EBA**. The reason is that the value of Δ is directly embedded in decision thresholds of **SEC1**, **SEC2** and **Non-adaptive Unif-EBA**, which is not the case for **SC-UCB**.

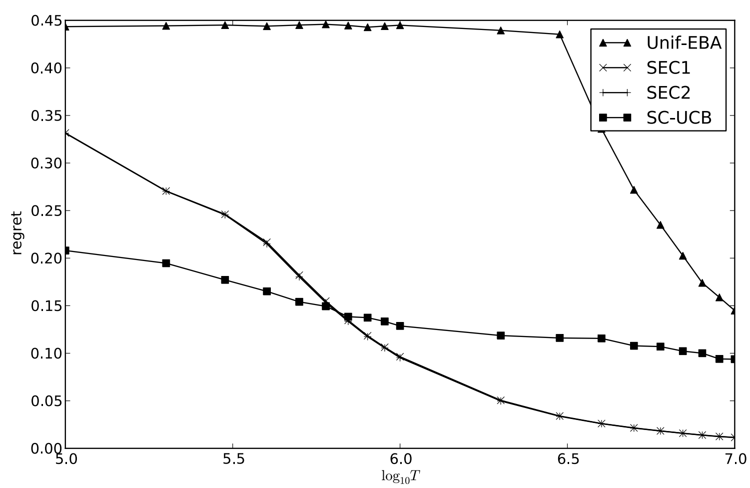

Figure 2: Numerical performances where $K = 20, \gamma = 0.75$, and $\Delta = 0.01$

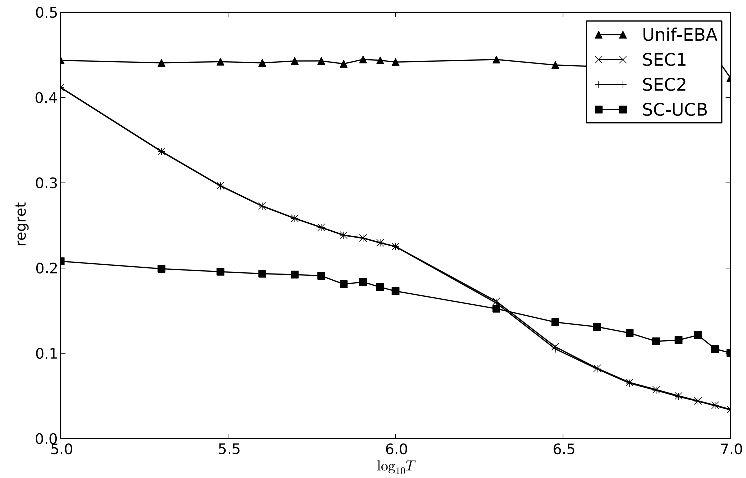

Figure 3: Numerical performances where $K = 20, \gamma = 0.75$, and $\Delta = 0.005$

We then investigate the effect of changing γ. Figure 4 shows the result for $K = 20, \gamma = 0.9$, and

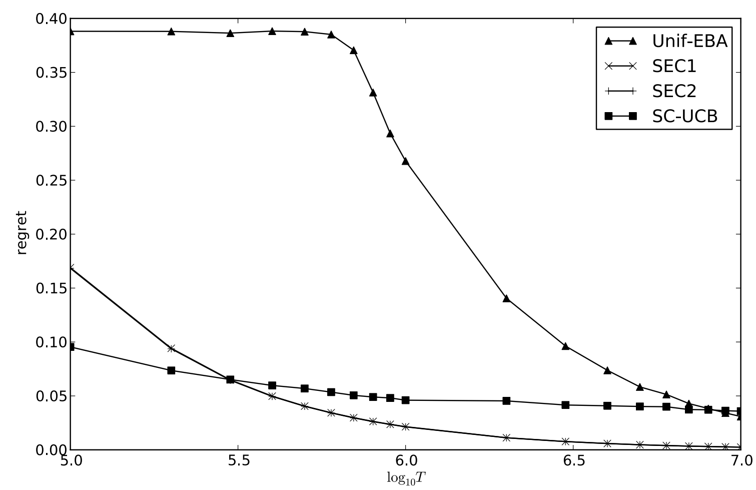

Figure 4: Numerical performances where $K = 20, \gamma = 0.9$, and $\Delta = 0.02$

$\Delta = 0.02$. As expected, the regrets of all algorithms in this case are smaller than the ones in the case of $\gamma = 0.75$ (Figure 1).

Finally, we keep $\gamma = 0.9, \Delta = 0.02$, and increase K. Figures 5 and 6 shows the results for $K = 50$ and $K = 100$, respectively. Again, we see that the increase of K affects all of the algorithms, but it has more effect on **Non-adaptive Unif-EBA**, **SEC1**, and **SEC2** than **SC-UCB**.

References

[1] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine Learning Journal*, 47(2-3):235–256, 2002.

[2] P. Auer and R. Ortner. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

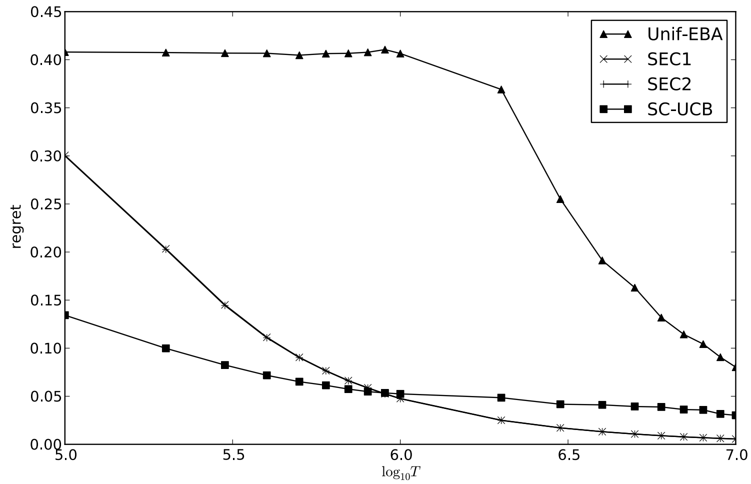

Figure 5: Numerical performances where $K = 50, \gamma = 0.90$, and $\Delta = 0.02$

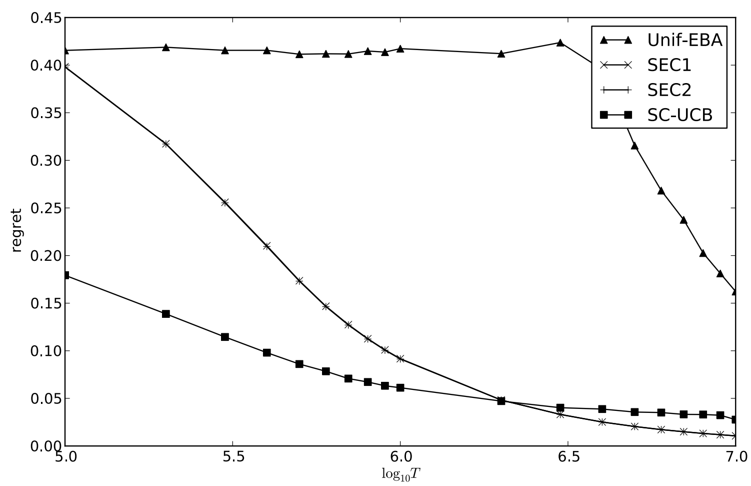

Figure 6: Numerical performances where $K = 100, \gamma = 0.90$, and $\Delta = 0.02$

[3] S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, 2011.

[4] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[5] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

[6] S. Mannor and J. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623–648, 2004.