1 Lower bound on regret for K-armed bernoulli bandits

For any horizon T and number of arms K, there exists a strategy for choosing rewards such that the expected regret for any algorithm is $\Omega\left(\sqrt{TK}\right)$. This strategy is oblivious to the algorithm and assigns rewards at random according to some distribution. Thus the regret bound applies to stochastic and adversarial bandits.

Theorem 1. *There exits a distribution over rewards such that:*

$$R(T) \geq \frac{1}{20}\min\{\sqrt{KT}, T\} \tag{1}$$

Proof. Consider a set of environments indexed by i. In environment i action i is the 'good' action. All other arms are slightly worse.

- In environment i, $r_{i,t} \sim Bernoulli(\frac{1}{2}+\epsilon)$ and $r_{j,t} \sim Bernoulli(\frac{1}{2})\ \ \forall j \neq i$
- $P_i\,(.)$ is the probability with respect to environment i
- $P_{unif}\,(.)$ is the probability with respect to an environment in which the expected reward for **all** arms is $\frac{1}{2}$
- $P_*\,(.)$ is the probability with respect to an environment sampled uniformly at random from $\{1...K\}$
- r_t is the reward received at time t
- $\boldsymbol{r}^t =< r_1, ..., r_t >$ is the history of rewards received upto time t
- A is the algorithm, maps $\boldsymbol{r}^{t-1} \rightarrow i_t$
- N_i is a random variable representing the number of times action i is selected by the algorithm.
- $G_A = \sum_{t=1}^{T} r_t$ is the total reward for the algorithm
- G_{max} is the total reward for playing the optimal arm in every timestep.

Lemma 2. *For any arm i,*

$$\mathbb{E}_i\left[N_i\right] - \mathbb{E}_{unif}\left[N_i\right] \leq \frac{T}{2}\sqrt{-\ln(1-4\epsilon^2)\mathbb{E}_{unif}\left[N_i\right]} \tag{2}$$

This says that the number of times we expect to play arm i in the environment in which it is optimal is not too much greater than the number of times we expect to play it if all arms are equal.

Proof.

$$\begin{aligned}
\mathbb{E}_i\left[N_i\right] - \mathbb{E}_{unif}\left[N_i\right] =& \sum_{\boldsymbol{r}\in\{0,1\}^T} N_i(\boldsymbol{r})\left(P_i\left(\boldsymbol{r}\right) - P_{unif}\left(\boldsymbol{r}\right)\right) && \leftarrow \text{definition of expectation} && (3)\\
\leq& \sum_{\boldsymbol{r}:P_i(\boldsymbol{r})\geq P_{unif}(\boldsymbol{r})} N_i(\boldsymbol{r})\left(P_i\left(\boldsymbol{r}\right) - P_{unif}\left(\boldsymbol{r}\right)\right) && \leftarrow \text{dropped only -ive terms} && (4)\\
\leq& T \sum_{\boldsymbol{r}:P_i(\boldsymbol{r})\geq P_{unif}(\boldsymbol{r})} \left(P_i\left(\boldsymbol{r}\right) - P_{unif}\left(\boldsymbol{r}\right)\right) && \leftarrow N_i \leq T\ \ \forall \boldsymbol{r} && (5)\\
=& \frac{T}{2}||P_i\left(\boldsymbol{r}\right) - P_{unif}\left(\boldsymbol{r}\right)||_1 && \leftarrow \text{see Thomas\&Cover 11.137} && (6)\\
\leq& \frac{T}{2}\sqrt{2\ln(2)KL\left(P_{unif}\left(\boldsymbol{r}\right)||P_i\left(\boldsymbol{r}\right)\right)} && \leftarrow \text{see Thomas\&Cover 11.138} && (7)\\
=& \frac{T}{2}\sqrt{-ln(2)lg\left(1-4\epsilon^2\right)\mathbb{E}_{unif}\left[N_i\right]} && \leftarrow \text{see section 1.1} && (8)\\
=& \frac{T}{2}\sqrt{-\ln(1-4\epsilon^2)\mathbb{E}_{unif}\left[N_i\right]} && \leftarrow \text{change of base} && (9)
\end{aligned}$$

□

Theorem 3. *For any algorithm A, if the distribution over rewards is selected unifirmly at random from environments $\{1...K\}$:*

$$E_*[G_{max} - G_A] \geq \epsilon \left(T - \frac{T}{K} - \frac{T}{2}\sqrt{-\frac{T}{K} ln(1-4\epsilon^2)} \right) \tag{10}$$

Proof.

$$\mathbb{E}_i\left[r_t\right] = \left(\frac{1}{2} + \epsilon\right) P_i\left(i_t = i\right) + \frac{1}{2}\left(1 - P_i\left(i_t = i\right)\right) \tag{11}$$

$$= \frac{1}{2} + \epsilon P_i\left(i_t = i\right) \tag{12}$$

$$\implies \mathbb{E}_i\left[G_A\right] = \sum_{t=1}^{T} \mathbb{E}_i\left[r_t\right] = \frac{T}{2} + \epsilon \mathbb{E}_i\left[N_i\right] \tag{13}$$

This gives us the expected gain given action i is the good action in terms of the number of times the algorithm A selects action i. The expected gain over all the environments i is:

$$\mathbb{E}_*\left[G_A\right] = \frac{1}{K}\sum_{i=1}^{K} \mathbb{E}_i\left[G_A\right] \tag{14}$$

$$= \frac{T}{2} + \frac{\epsilon}{K}\sum_{i=1}^{K} \mathbb{E}_i\left[N_i\right] \tag{15}$$

$$\mathbb{E}_*\left[G_{max}\right] = \left(\frac{1}{2} + \epsilon\right) T \tag{16}$$

From lemma 2

$$\sum_{i=1}^{K} \mathbb{E}_i\left[N_i\right] \leq \sum_{i=1}^{K} \left(\mathbb{E}_{unif}\left[N_i\right] + \frac{T}{2}\sqrt{-\ln(1-4\epsilon^2)\mathbb{E}_{unif}\left[N_i\right]} \right) \tag{17}$$

Now $\sum_{i=1}^{K} \mathbb{E}_{unif}\left[N_i\right] = \mathbb{E}_{unif}\left[\sum_{i=1}^{K} N_i\right] = \mathbb{E}_{unif}\left[T\right] = T$

Note: we cannot do the same with $\sum_{i=1}^{K} \mathbb{E}_i\left[N_i\right]$ as $\mathbb{E}_i\left[.\right]$ is with respect to a different distribution for each i.

$$\implies \sum_{i=1}^{K} \mathbb{E}_i\left[N_i\right] \leq T + \frac{T}{2}\sqrt{-\ln(1-4\epsilon^2)KT} \qquad \leftarrow \text{via Jenson's Inequality} \tag{18}$$

$$\implies \mathbb{E}_*\left[G_A\right] \leq \frac{T}{2} + \epsilon\left(\frac{T}{K} + \frac{T}{2}\sqrt{-\frac{T}{K}\ln(1-4\epsilon^2)}\right) \tag{19}$$

$$\implies \mathbb{E}_*\left[G_{max} - G_A\right] \geq \epsilon\left(T - \frac{T}{K} - \frac{T}{2}\sqrt{-\frac{T}{K} ln(1-4\epsilon^2)}\right) \tag{20}$$

□

□

1.1 Calculation of KL divergence

$$KL\left(p(x_1,...,x_n)||q(x_1,...,x_n)\right) = \sum_{i=1}^{N} KL\left(p(x_i|\boldsymbol{x}^{i-1})||q(x_i|\boldsymbol{x}^{i-1})\right) \qquad \leftarrow \text{chain rule for KL divergence} \tag{21}$$

$$\text{where } KL\left(p(x_i|\boldsymbol{x}^{i-1})||q(x_i|\boldsymbol{x}^{i-1})\right) = \sum_{\boldsymbol{x}^{i-1}} p(\boldsymbol{x}^{i-1}) \sum_{x_i} p(x_i|\boldsymbol{x}^{i-1}) lg\left(\frac{p(x_i|\boldsymbol{x}^{i-1})}{q(x_i|\boldsymbol{x}^{i-1})}\right) \tag{22}$$

So

$$KL\left(P_{unif}||P_i\right) = \sum_{t=1}^{T} \left(\underbrace{\sum_{\boldsymbol{r}^{t-1}} P_{unif}\left(\boldsymbol{r}^{t-1}\right)}_{\text{expectation over history}} \underbrace{\sum_{r_t \in \{0,1\}} P_{unif}\left(r_t|\boldsymbol{r}^{t-1}\right) lg\left(\frac{P_{unif}\left(r_t|\boldsymbol{r}^{t-1}\right)}{P_i\left(r_t|\boldsymbol{r}^{t-1}\right)}\right)}_{\text{KL divergence at time t}} \right) \tag{23}$$

Now

$$P_{unif}\left(r_t|\boldsymbol{r}^{t-1}\right) = \frac{1}{2} \ \ \forall\, r_t, \boldsymbol{r}^{t-1} \tag{24}$$

$$P_i\left(r_t|\boldsymbol{r}^{t-1}\right) = \begin{cases} (\frac{1}{2}+\epsilon)^{r_t}(\frac{1}{2}-\epsilon)^{1-r_t} & \text{if } A(\boldsymbol{r}^{t-1}) = i \\ \frac{1}{2} & \text{otherwise} \end{cases} \tag{25}$$

Let B be the set of histories that lead the algorithm to select the good arm, $B = \left\{\boldsymbol{r}^{t-1} : A(\boldsymbol{r}^{t-1}) = i\right\}$. Note: $\boldsymbol{r}^{t-1}$ is sufficient to determine i_t for a deterministic algorithm A in the bandit setting. At the first timestep A will select $i_1(A)$, it then receives reward r_1, selects $i_2(A, r_1)$ and so on.

$$KL\left(P_{unif}||P_i\right) = \sum_{t=1}^{T} \left(\sum_{B} P_{unif}\left(\boldsymbol{r}^{t-1}\right) KL\left(\frac{1}{2}||\frac{1}{2}+\epsilon\right) + \sum_{B^c} P_{unif}\left(\boldsymbol{r}^{t-1}\right) KL\left(\frac{1}{2}||\frac{1}{2}\right) \right) \tag{26}$$

$$= KL\left(\frac{1}{2}||\frac{1}{2}+\epsilon\right) \sum_{t=1}^{T} \left(\sum_{B} P_{unif}\left(\boldsymbol{r}^{t-1}\right) \right) \qquad \leftarrow \text{ as } kl(\frac{1}{2}||\frac{1}{2}) = 0 \tag{27}$$

$$= KL\left(\frac{1}{2}||\frac{1}{2}+\epsilon\right) \sum_{t=1}^{T} \left(P_{unif}\left(i_t = i\right)\right) \tag{28}$$

$$= KL\left(\frac{1}{2}||\frac{1}{2}+\epsilon\right) \mathbb{E}_{unif}\left[N_i\right] \tag{29}$$

$$= -\frac{1}{2} lg\left(1 - 4\epsilon^2\right) \mathbb{E}_{unif}\left[N_i\right] \tag{30}$$

1.2 Jenson's inequality

Jenson's inequality states that for a concave function ϕ:

$$\frac{\sum_{i=1}^{K}\phi(x_i)}{K} \le \phi\left(\frac{\sum_{i=1}^{K}x_i}{K}\right) \tag{31}$$

$$\implies \sum_{i=1}^{K}\sqrt{x_i} \le \sqrt{K\sum_{i=1}^{K}x_i} \tag{32}$$

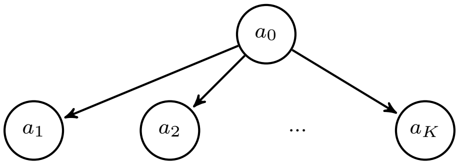
$$\implies \sum_{i=1}^{K}\sqrt{\mathbb{E}_{unif}\left[N_i\right]} \le \sqrt{KT} \tag{33}$$

$$\implies \sum_{i=1}^{K}\mathbb{E}_i\left[N_i\right] \le T + \frac{T}{2}\sqrt{-\ln(1-4\epsilon^2)KT} \tag{34}$$

2 Lower bound for specific feedback graph

Now we move to considering a Bernoulli bandit problem with the feedback graph shown in figure 1. The environments specifying rewards over arms $\{1...K\}$ are defined identically to the standard bandit case. The expected reward for a_0 (in all environments) is 0 (this is the worst case).

Figure 1: Feedback graph



As for the standard bandit case,

$$\mathbb{E}_i\left[N_i\right] - \mathbb{E}_{unif}\left[N_i\right] \le \frac{T}{2}\sqrt{2\ln(2)KL\left(P_{unif}\left(\boldsymbol{h}\right)||P_i\left(\boldsymbol{h}\right)\right)} \tag{35}$$

The only difference is that the sequence of rewards $\boldsymbol{r}$ is no longer sufficient to determine N_i because when we select a_0 we get feedback on all the other arms that does not contribute to the reward but can be leveraged by an algorithm. Therefore the KL divergence is now over distributions over the history $\boldsymbol{h}$ that includes this additional information.

2.1 Calculation of KL divergence

Define

$$h_t = \begin{cases}(r_t, o_{1t}, ..., o_{Kt}) & \text{if } i = 0\\ r_t & \text{otherwise}\end{cases} \tag{36}$$

Where o_{jt} is the observed reward for arm j at timestep t.

$$KL\left(P_{unif}\left(\boldsymbol{h}\right)||P_i\left(\boldsymbol{h}\right)\right) = \sum_{t=1}^{T} KL\left(P_{unif}\left(h_t|\boldsymbol{h}^{t-1}\right)||P_i\left(h_t|\boldsymbol{h}^{t-1}\right)\right) \tag{37}$$

At each time step t, $i_t = A(\boldsymbol{h}^{t-1})$ There are three cases for the value of the KL divergence.

1. The algorithm selects the optimal action. $i_t = i$, (as for the standard bandit case)

$$P_{unif}\left(h_t|\boldsymbol{h}^{t-1}\right) = P_{unif}\left(r_t|\boldsymbol{h}^{t-1}\right) = \frac{1}{2} \tag{38}$$

$$P_i\left(h_t|\boldsymbol{h}^{t-1}\right) = P_i\left(r_t|\boldsymbol{h}^{t-1}\right) = (\frac{1}{2}+\epsilon)^{r_t}(\frac{1}{2}-\epsilon)^{1-r_t} \tag{39}$$

$$\implies kl(\frac{1}{2}||\frac{1}{2}+\epsilon) \tag{40}$$

2. The algorithm selects the revealing action. $i_t = 0$

$$P_{unif}\left(h_t|\boldsymbol{h}^{t-1}\right) = B_0(r_t)\prod_{j\in\{1\ldots K\}} B_{\frac{1}{2}}(o_j) \qquad \text{where} B_\alpha(x) \implies x \sim Bernoulli(\alpha) \tag{41}$$

$$P_i\left(h_t|\boldsymbol{h}^{t-1}\right) = B_0(r_t)\prod_{j\in\{1\ldots K\}/i} B_{\frac{1}{2}}(o_j)B_{\frac{1}{2}+\epsilon}(o_i) \tag{42}$$

$$\implies kl(\frac{1}{2}||\frac{1}{2}+\epsilon) \tag{43}$$

Since the only term that differs P_{unif} and P_i is that for o_i and,

$$KL\left(\prod_{i=1}^{n} P_i(x_i)||\prod_{i=1}^{n} Q_i(x_i)\right) = \sum_{i=1}^{n} KL\left(P_i(x_i)||Q_i(x_i)\right) \tag{44}$$

3. Otherwise. $i_t \in \{1,..,K\}/i$, (as for the standard bandit case)

$$P_{unif}\left(h_t|\boldsymbol{h}^{t-1}\right) = P_i\left(h_t|\boldsymbol{h}^{t-1}\right) = \frac{1}{2} \tag{45}$$

$$\implies kl(\frac{1}{2}||\frac{1}{2}) = 0 \tag{46}$$

We divide the histories into corresponding sets. $S_1 = \left\{\boldsymbol{h}^{t-1} : A(\boldsymbol{h}^{t-1}) = i\right\}$, $S_2 = \left\{\boldsymbol{h}^{t-1} : A(\boldsymbol{h}^{t-1}) = 0\right\}$,$S_3 = \left\{\boldsymbol{h}^{t-1} : A(\boldsymbol{h}^{t-1}) \in \{1,..,K\}/i\right\}$

$$KL\left(P_{unif}||P_i\right) = \sum_{t=1}^{T}\left(\sum_{S_1} P_{unif}\left(\boldsymbol{h}^{t-1}\right) kl(\frac{1}{2}||\frac{1}{2}+\epsilon) + \sum_{S_2} P_{unif}\left(\boldsymbol{h}^{t-1}\right) kl(\frac{1}{2}||\frac{1}{2}+\epsilon) + \sum_{S_3} P_{unif}\left(\boldsymbol{h}^{t-1}\right) kl(\frac{1}{2}||\frac{1}{2})\right) \tag{47}$$

$$= kl(\frac{1}{2}||\frac{1}{2}+\epsilon)\sum_{t=1}^{T}\left(\sum_{S_1} P_{unif}\left(\boldsymbol{h}^{t-1}\right) + \sum_{S_2} P_{unif}\left(\boldsymbol{h}^{t-1}\right)\right) \tag{48}$$

$$= kl(\frac{1}{2}||\frac{1}{2}+\epsilon)\sum_{t=1}^{T}\left(P_{unif}\left(i_t = i\right) + P_{unif}\left(i_t = 0\right)\right) \tag{49}$$

$$= kl(\frac{1}{2}||\frac{1}{2}+\epsilon)\left(\mathbb{E}_{unif}\left[N_i\right] + \mathbb{E}_{unif}\left[N_0\right]\right) \tag{50}$$

$$= -\frac{1}{2}lg(1-4\epsilon^2)\left(\mathbb{E}_{unif}\left[N_i\right] + \mathbb{E}_{unif}\left[N_0\right]\right) \tag{51}$$

Similarly (I think)

$$KL\left(P_i||P_{unif}\right) = kl(\frac{1}{2}+\epsilon||\frac{1}{2})\left(\mathbb{E}_i\left[N_i\right] + \mathbb{E}_i\left[N_0\right]\right) \tag{52}$$

$$= (\frac{1}{2}lg((1+2\epsilon)(1-2\epsilon)) + \epsilon lg\left(\frac{1+2\epsilon}{1-2\epsilon}\right))\left(\mathbb{E}_i\left[N_i\right] + \mathbb{E}_i\left[N_0\right]\right) \tag{53}$$

$$\leq -\frac{1}{2}lg(1-4\epsilon^2)\left(\mathbb{E}_i\left[N_i\right] + \mathbb{E}_i\left[N_0\right]\right) \tag{54}$$

2.2 Bounding regret

We will consider only algorithms that select N_0 at most ϵT times. Once an algorithm has selected N_0 ϵT times, we will truncate it and play any other action at random. The regret for the non-truncated algorithm is at least $\epsilon T(\frac{1}{2}+\epsilon) \geq \frac{\epsilon T}{2}$. Truncating adds at most $(T - \epsilon T)\epsilon \leq \epsilon T$. Thus, the assumption that $N_0 \leq \epsilon T$ increases the regret by at most a constant factor of 3.

$$\mathbb{E}_i\left[r_t\right] = (\frac{1}{2}+\epsilon)P_i\left(i_t = i\right) + 0P_i\left(i_t = 0\right) + \frac{1}{2}P_i\left(i_t \in \left\{1...K\right\}/i\right) \tag{55}$$

$$= (\frac{1}{2}+\epsilon)P_i\left(i_t = i\right) + \frac{1}{2}(1 - P_i\left(i_t = i\right) - P_i\left(i_t = 0\right)) \tag{56}$$

$$= \frac{1}{2} + \epsilon P_i\left(i_t = i\right) - \frac{1}{2}P_i\left(i_t = 0\right) \tag{57}$$

$$\mathbb{E}_i\left[G_A\right] = \sum_{t=1}^{T}\mathbb{E}_i\left[r_t\right] = \frac{T}{2} + \epsilon\mathbb{E}_i\left[N_i\right] - \frac{1}{2}\mathbb{E}_i\left[N_0\right] \tag{58}$$

$$\mathbb{E}_*\left[G_A\right] = \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_i\left[G_A\right] = \frac{T}{2} + \frac{\epsilon}{K}\sum_{i=1}^{K}\mathbb{E}_i\left[N_i\right] - \frac{1}{2K}\sum_{i=1}^{K}\mathbb{E}_i\left[N_0\right] \tag{59}$$

$$\mathbb{E}_*\left[G_{max}\right] = (\frac{1}{2}+\epsilon)T \tag{60}$$

$$\implies R(T) =\epsilon\left(T - \frac{1}{K}\sum_{i=1}^{K} \mathbb{E}_i\left[N_i\right]\right) + \frac{1}{2K}\sum_{i=1}^{K} \mathbb{E}_i\left[N_0\right] \tag{61}$$

$$\geq\epsilon\left(T-\frac{T}{K}-\frac{T}{2}\sqrt{-\frac{1}{K}ln(1-4\epsilon^2)(T+\sum_{i=1}^{K}\mathbb{E}_{unif}\left[N_0\right])}\right)+\frac{1}{2K}\sum_{i=1}^{K}\mathbb{E}_i\left[N_0\right] \tag{62}$$

$$=\epsilon\left(T-\frac{T}{K}-\frac{T}{2}\sqrt{-ln(1-4\epsilon^2)(\frac{T}{K}+\mathbb{E}_{unif}\left[N_0\right])}\right)+\frac{1}{2}\mathbb{E}_*\left[N_0\right] \tag{63}$$

$$\geq\epsilon\left(T-\frac{T}{K}-T\epsilon\sqrt{2(\frac{T}{K}+\mathbb{E}_{unif}\left[N_0\right])}\right)+\frac{1}{2}\mathbb{E}_*\left[N_0\right] \qquad \text{if } \epsilon < \frac{2}{5} \tag{64}$$

$$\geq\frac{\epsilon}{3}\left(T-\frac{T}{K}-T\epsilon\sqrt{2T(\frac{1}{K}+\epsilon)}\right) \qquad \text{assuming } \mathbb{E}_{unif}\left[N_0\right] \leq \epsilon T \tag{65}$$

Let

$$\epsilon = \begin{cases} \frac{1}{4}\sqrt{\frac{K}{T}} & \text{if } K < (16T)^{1/3} \\ (16T)^{-1/3} & \text{otherwise} \end{cases} \tag{66}$$

If $K < (16T)^{1/3}$

$$R(T) \geq\frac{1}{3}\left(\frac{1}{4}\sqrt{KT}-\frac{1}{4}\sqrt{\frac{T}{K}}-\frac{1}{16\sqrt{2}}\sqrt{4KT+K^{5/2}T^{1/2}}\right) \tag{67}$$

$$\geq\frac{1}{3}\left(\frac{1}{4}\sqrt{KT}-\frac{1}{4}\sqrt{\frac{T}{K}}-\frac{1}{16\sqrt{2}}\sqrt{8KT}\right) \tag{68}$$

$$=\frac{1}{3}\left(\frac{1}{8}\sqrt{KT}-\frac{1}{4}\sqrt{\frac{T}{K}}\right) \tag{69}$$

$$\geq\frac{1}{72}\sqrt{KT} \qquad \text{if } K > 2 \tag{70}$$

If $K \geq (16T)^{1/3}$

$$R(T) \geq\frac{T^{2/3}}{3(16^{1/3})}\left(1-\frac{1}{K}-\frac{1}{2(2^{5/6})}\sqrt{\frac{4T^{1/3}}{K}+2^{2/3}}\right) \tag{71}$$

$$\geq\frac{T^{2/3}}{3(16^{1/3})}\left(1-\frac{1}{K}-\frac{1}{2(2^{5/6})}\sqrt{2(2^{2/3})}\right) \tag{72}$$

$$=\frac{T^{2/3}}{3(16^{1/3})}\left(1-\frac{1}{K}-\frac{1}{2}\right) \tag{73}$$

$$\geq\frac{1}{46}T^{2/3} \qquad \text{if } K > 2 \tag{74}$$

$$\tag{75}$$

3 Graph changes with time

Same as before we have a single revealing action. However, now at each timestep it reveals the rewards for a half the arms at random according to the prior probability vector $\boldsymbol{q}$

- Assume that $q_j \in [0, \frac{1}{2}]$ and $q_1 \leq q_2, \leq ... \leq, q_N$
- Let $m \in \{2...N\} = min_i : q_j \geq \frac{1}{j}$
- Let $M = \{j : j \leq m\}$, (the set of unbalanced arms).
- let D_t be the set of arms revealed by the revealing action at timestep t

3.1 KL divergence

As for previous section except now if the revealing action is selected, $P_{unif}\left(h_t|\boldsymbol{h}^{t-1}\right) = P_i\left(h_t|\boldsymbol{h}^{t-1}\right)$ unless $i \in D_t$

The probability $P(i \in D_t)$ depends only on $\boldsymbol{q}$ (and is independent of the history at anytime step $\boldsymbol{h}$).

So:

$$KL\left(P_{unif}||P_i\right) = kl(\frac{1}{2}||\frac{1}{2}+\epsilon)\left(\mathbb{E}_{unif}\left[N_i\right] + \sum_{t=1}^{T} P_{unif}\left(i_t = 0\right) P(i \in D_t)\right) \tag{76}$$

$$= kl(\frac{1}{2}||\frac{1}{2}+\epsilon)\left(\mathbb{E}_{unif}\left[N_i\right] + q_i \mathbb{E}_{unif}\left[N_0\right]\right) \tag{77}$$

$$\implies \mathbb{E}_i\left[N_i\right] \leq \mathbb{E}_{unif}\left[N_i\right] + \sqrt{2}T\epsilon\sqrt{\mathbb{E}_{unif}\left[N_i\right] + q_i\mathbb{E}_{unif}\left[N_0\right]} \qquad \text{assuming } \epsilon < \frac{2}{5} \tag{78}$$

Let $\mathbb{E}_*\left[N_i\right] = \sum_{i=1}^{K} w_i \mathbb{E}_i\left[N_i\right]$

$$\mathbb{E}_*\left[N_i\right] \leq \sum_i w_i \mathbb{E}_{unif}\left[N_i\right] + \sqrt{2}T\epsilon\sqrt{\sum_i w_i\mathbb{E}_{unif}\left[N_i\right] + \mathbb{E}_{unif}\left[N_0\right]\sum_i w_i q_i} \tag{79}$$

Choose the weights such that $P(i \in M) = \frac{1}{2}$ ie, there is a 50% chance of choosing an environment in which the optimal arm is in the set M.

$$w_i = \begin{cases} \frac{1}{2(m-1)} & \text{if } i \leq m \\ \frac{1}{2(K-m+1)} & \text{otherwise} \end{cases} \tag{80}$$

Now $\sum_i w_i \mathbb{E}_{unif}\left[N_i\right] = \mathbb{E}_{unif}\left[\frac{1}{2(m-1)}\sum_{i=1}^{m-1} N_i + \frac{1}{2(K-m+1)}\sum_{i=m}^{K} N_i\right] \leq \frac{T}{2(m-1)}$, as $m - 1 < K - m + 1$, since for every unbalanced arm in m its partner is balanced.

$$\mathbb{E}_*\left[N_i\right] \leq \frac{T}{2(m-1)} + \sqrt{2}T\epsilon\sqrt{\frac{T}{2(m-1)} + \mathbb{E}_{unif}\left[N_0\right]\sum_i w_i q_i} \tag{81}$$

$$\sum_i w_i q_i \leq \frac{1}{2(m-1)}\sum_{i<m}\frac{1}{m} + \frac{1}{2(k-m+1)}\sum_{i\geq m} 1 = \frac{1}{2}\left(\frac{1}{m}+1\right) \tag{82}$$

3.2 Bounding regret

We can now assume that the algorithm doesn't select N_0 more than $\epsilon T(1-\frac{1}{2(m-1)})$ times, since we can truncate all algorithms at that point and then play arms from $i \in M$ at random.

$$\begin{aligned}
R(T) &= \epsilon\left(T - \mathbb{E}_*\left[N_i\right]\right) + \frac{1}{2}\mathbb{E}_*\left[N_0\right] && (83)\\
&\geq \epsilon\left(T - \frac{T}{2(m-1)} - \sqrt{2}T\epsilon\sqrt{\frac{T}{2(m-1)} + \frac{1}{2}\left(\frac{1}{m}+1\right)\mathbb{E}_{unif}\left[N_0\right]}\right) && (84)\\
&\geq \frac{\epsilon}{3}\left(T - \frac{T}{2(m-1)} - \sqrt{2}T\epsilon\sqrt{\frac{T}{2(m-1)} + \frac{1}{2}\left(\frac{1}{m}+1\right)\epsilon T(1-\frac{1}{2(m-1)})}\right) && (85)\\
&= \frac{\epsilon}{3}\left(T - \frac{T}{2(m-1)} - \sqrt{2}T\epsilon\sqrt{\frac{T}{2(m-1)} + \epsilon T\frac{m(2m-1)-3}{4m(m-1)}}\right) && (86)\\
&\sim \frac{\epsilon}{3}\left(T - \frac{T}{2(m-1)} - \sqrt{2}T\epsilon\sqrt{\frac{T}{2(m-1)} + \frac{\epsilon T}{2}}\right) \qquad \frac{m(2m-1)-3}{4m(m-1)} \leq \frac{9}{17} \sim \frac{1}{2} && (87)
\end{aligned}$$

No dependence on K this just cannot work.

$$\begin{aligned}
R(T) &= \epsilon\left(T - \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_i\left[N_i\right]\right) + \frac{1}{2K}\sum_{i=1}^{K}\mathbb{E}_i\left[N_0\right] \qquad \leftarrow \text{as before} && (88)\\
&\geq \epsilon\left(T - \frac{T}{K} - T\epsilon\sqrt{\frac{2}{K}\left(\sum_{i}\mathbb{E}_{unif}\left[N_i\right] + \sum_{i\in M}\frac{1}{m}\mathbb{E}_{unif}\left[N_0\right] + \sum_{i\notin M}\mathbb{E}_{unif}\left[N_0\right]\right)}\right) + \frac{1}{2}\mathbb{E}_*\left[N_0\right] && (89)\\
&= \epsilon\left(T - \frac{T}{K} - T\epsilon\sqrt{2\left(\frac{T}{K} + \frac{1}{m}\mathbb{E}_{unif}\left[N_0\right]\right)}\right) + \frac{1}{2}\mathbb{E}_*\left[N_0\right] && (90)\\
&\geq \frac{\epsilon}{3}\left(T - \frac{T}{K} - T\epsilon\sqrt{2T\left(\frac{1}{K} + \frac{\epsilon}{m}\right)}\right) \qquad \text{assuming } \mathbb{E}_{unif}\left[N_0\right] \leq \epsilon T && (91)
\end{aligned}$$

$$\begin{aligned}
R(T) &= \epsilon\left(T - \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_i\left[N_i\right]\right) + \frac{1}{2K}\sum_{i=1}^{K}\mathbb{E}_i\left[N_0\right] \qquad \leftarrow \text{as before} && (92)\\
&\leq \epsilon\left(T - \frac{T}{K} - T\epsilon\sqrt{\frac{2}{K}\left(\sum_{i\in M}\mathbb{E}_{unif}\left[N_i\right] + \sum_{i\notin M}\left(\mathbb{E}_{unif}\left[N_i\right] + \frac{1}{m}\mathbb{E}_{unif}\left[N_0\right]\right)\right)}\right) + \frac{1}{2}\mathbb{E}_*\left[N_0\right] && (93)\\
&= \epsilon\left(T - \frac{T}{K} - T\epsilon\sqrt{2\left(\frac{T}{K} + \frac{1}{m}\mathbb{E}_{unif}\left[N_0\right]\right)}\right) + \frac{1}{2}\mathbb{E}_*\left[N_0\right] && (94)\\
&\geq \frac{\epsilon}{3}\left(T - \frac{T}{K} - T\epsilon\sqrt{2T\left(\frac{1}{K} + \frac{\epsilon}{m}\right)}\right) \qquad \text{assuming } \mathbb{E}_{unif}\left[N_0\right] \leq \epsilon T && (95)
\end{aligned}$$

Let:

$$\epsilon = \begin{cases} \frac{1}{4}\sqrt{\frac{K}{T}} & \text{if } K > m^{2/3}(16T)^{1/3} \\ m^{1/3}(16T)^{-1/3} & \text{otherwise} \end{cases} \tag{96}$$

If $K < m^{2/3}(16T)^{1/3}$

$$\begin{aligned}
R(T) &\geq \frac{\sqrt{KT}}{12}\left(\frac{1}{2} - \frac{1}{K}\right) && (97)\\
&\geq \frac{1}{72}\sqrt{KT} \qquad \text{if } K > 2 && (98)
\end{aligned}$$

If $K \geq m^{2/3}(16T)^{1/3}$

$$\begin{aligned}
R(T) &\geq \frac{m^{1/3}T^{2/3}}{3(16^{1/3})}\left(\frac{1}{2} - \frac{1}{K}\right) && (99)\\
&\geq \frac{1}{46}m^{1/3}T^{2/3} \qquad \text{if } K > 2 && (100)
\end{aligned}$$

3.3 Previous Attempt at bounding regret

If I just let $\mathbb{E}_{unif}\left[N_0\right] = T$ and $\mathbb{E}_*\left[N_0\right] = 0$, I get $R \sim \Omega\left(\sqrt{T}\right)$ - so not tight enough.

By similar appeal to KL divergence, we get

$$\mathbb{E}_i\left[N_0\right] \geq \mathbb{E}_{unif}\left[N_0\right] - \frac{T}{2}\sqrt{-ln(1-4\epsilon^2)(\mathbb{E}_{unif}\left[N_i\right] + \mathbb{E}_{unif}\left[N_0\right])} \tag{101}$$

$$\implies \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_i\left[N_0\right] \geq \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{unif}\left[N_0\right] - \frac{T}{2K}\sum_{i=1}^{K}\sqrt{-ln(1-4\epsilon^2)(\mathbb{E}_{unif}\left[N_i\right] + \mathbb{E}_{unif}\left[N_0\right])} \tag{102}$$

$$\implies \mathbb{E}_*\left[N_0\right] \geq \mathbb{E}_{unif}\left[N_0\right] - \frac{T}{2}\sqrt{-ln(1-4\epsilon^2)(\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{unif}\left[N_i\right] + \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{unif}\left[N_0\right])} \tag{103}$$

$$= \mathbb{E}_{unif}\left[N_0\right] - \frac{T}{2}\sqrt{-ln(1-4\epsilon^2)(\frac{T}{K} + \mathbb{E}_{unif}\left[N_0\right])} \tag{104}$$

$$\implies R(T) \geq \epsilon(T - \frac{T}{K}) - (\epsilon + \frac{1}{2})\frac{T}{2}\sqrt{-ln(1-4\epsilon^2)(\frac{T}{K} + \mathbb{E}_{unif}\left[N_0\right])} + \frac{1}{2}\mathbb{E}_{unif}\left[N_0\right] \tag{105}$$

For comparison, if we assumed $\mathbb{E}_{unif}\left[N_0\right] = \mathbb{E}_*\left[N_0\right]$ in equation 63 we would have:

$$R(T) \geq \epsilon(T - \frac{T}{K}) - \epsilon\frac{T}{2}\sqrt{-ln(1-4\epsilon^2)(\frac{T}{K} + \mathbb{E}_{unif}\left[N_0\right])} + \frac{1}{2}\mathbb{E}_{unif}\left[N_0\right] \tag{106}$$

So the regret bound is pulled down by an extra $\frac{T}{4}\sqrt{-ln(1-4\epsilon^2)(\frac{T}{K} + \mathbb{E}_{unif}\left[N_0\right])}$

Argument that we cannot get a regret bound starting from equation 105

Let:

$$R_1 = \epsilon(T - \frac{T}{K}) - (\epsilon + \frac{1}{2})\frac{T}{2}\sqrt{-ln(1-4\epsilon^2)(\frac{T}{K} + \mathbb{E}_{unif}\left[N_0\right])} + \frac{1}{2}\mathbb{E}_{unif}\left[N_0\right] \tag{107}$$

such that $R(T) \geq R_1$

$$
\begin{aligned}
R_1 &\leq \epsilon T - (\epsilon + \frac{1}{2})\frac{T}{2}\sqrt{-ln(1-4\epsilon^2)(\frac{T}{K} + \mathbb{E}_{unif}\left[N_0\right])} + \frac{1}{2}\mathbb{E}_{unif}\left[N_0\right] && (108)\\
&\leq \epsilon T - \frac{1}{2}\frac{T}{2}\sqrt{-ln(1-4\epsilon^2)(\frac{T}{K} + \mathbb{E}_{unif}\left[N_0\right])} + \frac{1}{2}\mathbb{E}_{unif}\left[N_0\right] && (109)\\
&\leq \epsilon T - \frac{T}{4}\sqrt{4\epsilon^2(\frac{T}{K} + \mathbb{E}_{unif}\left[N_0\right])} + \frac{1}{2}\mathbb{E}_{unif}\left[N_0\right] && (110)\\
&\leq \epsilon T - \epsilon\frac{T}{2}\sqrt{\mathbb{E}_{unif}\left[N_0\right]} + \frac{1}{2}\mathbb{E}_{unif}\left[N_0\right] && (111)\\
&\leq \frac{1}{2}\left(T - \frac{T}{2}\sqrt{\mathbb{E}_{unif}\left[N_0\right]}\right) + \frac{1}{2}\mathbb{E}_{unif}\left[N_0\right] && (112)\\
&= \frac{1}{2}\left(2T - \frac{T^{3/2}}{2}\right) \qquad \text{if } \mathbb{E}_{unif}\left[N_0\right] = T && (113)\\
&\leq 0 \qquad \text{if } T > 16 && (114)
\end{aligned}
$$