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 $r_{j,t} \sim Bernoulli(\frac{1}{2})\ \ \forall j \neq i$
- $P_i\left(.\right)$ is the probability with respect to environment i
- $P_{unif}\left(.\right)$ is the probability with respect to an environment in which the expected reward for **all** arms is $\frac{1}{2}$
- $P_*\left(.\right)$ 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 $\sum_{i=1}^{K} \mathbb{E}_{unif}\left[N_i\right] = T$ (because $\sum_{i=1}^{K} N_i = T$? but doesn't that imply $\sum_{i=1}^{K} \mathbb{E}_i\left[N_i\right] = T$?)

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

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

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

□

□

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, $B = \left\{\boldsymbol{r}^{t-1} : A(\boldsymbol{r}^{t-1}) = i\right\}$

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

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

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

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

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

1.2 Jenson's inequality

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

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

$$\implies \sum_{i=1}^{K}\sqrt{\mathbb{E}_{unif}\left[N_i\right]} \leq \sqrt{KT} \tag{32}$$

$$\implies \sum_{i=1}^{K}\mathbb{E}_i\left[N_i\right] \leq T + \frac{T}{2}\sqrt{-\ln(1-4\epsilon^2)KT} \tag{33}$$