Multi-Armed Bandits 3

Ashwinee Panda, David Vendrow, William Song

October 2018

1 Preliminaries

$T :=$ total number of timesteps for which we are pulling an arm
$X_{i,t} :=$ reward for pulling arm i on round t.
$T_i(t) :=$ number of times arm i is pulled before round t

The two-armed stochastic bandit is given by:
$(X_{1,1}, X_{1,2} \dots X_{1,T})$ i.i.d. $\sim Ber(\mu_1)$,
$(X_{1,1}, X_1 2, 1 \dots X_{T,1})$ i.i.d. $\sim Ber(\mu_2)$

$\Delta := \mu_{best} - \mu_{worse}$ is the gap between the two arms

The goal is to minimize "pseudo-regret" given by:

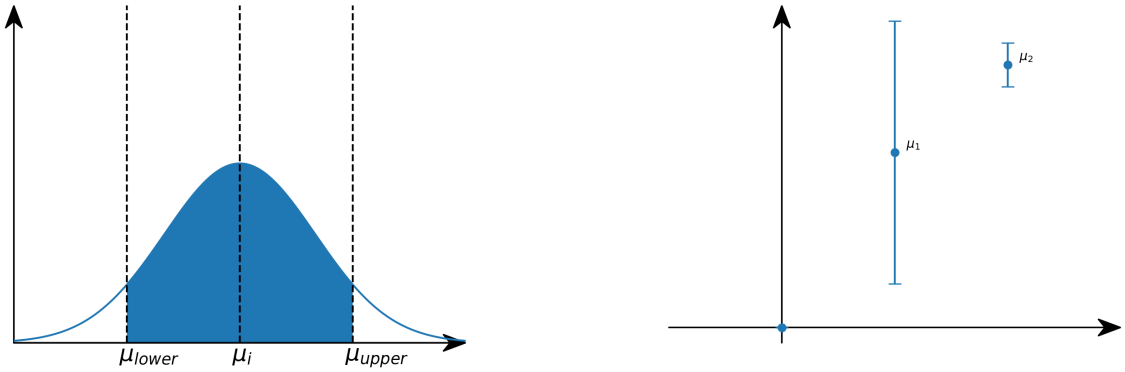
$$\overline{R_T} := T\mu^* - \sum_{t=1}^{T} \mathbb{E}[\mu_{I_t}] = \Delta\, \mathbb{E}[T_2(T)]$$

where $\sum_{t=1}^{T} \mathbb{E}[\mu_{I_t}]$ is the expected total reward our strategy achieves and $T\mu^*$ is the best expected total reward in hindsight, achieved by picking the best arm at each time step.

2 Elegant UCB Recap

Without loss of generality, assume that arm 1 is better than arm 2, so we can write $\Delta = \mu_1 - \mu_2$. The goal is to minimize pseudo-regret (defined above), which is a measure of how our performance compares to how well we could have done if we knew μ_1 and μ_2 beforehand.

Elegant UCB admits one parameter, $\delta_t \sim t^{-\alpha}$, which controls how large we want our confidence intervals to be. A larger confidence interval signifies that we expect the "truth" to be within a larger margin of error around the sample mean; therefore, a larger confidence interval is less precise.



(a) Confidence Interval

(b) UCB would choose μ_1 in this example

Figure 1

More formally: $\Pr[\hat{\mu}_2[\mu_{low}, \mu_{high}]] \geq 1 - \delta_t$

The action (which arm to pull) is selected according to the following rule:

$$I_t = \arg\max_{i \in \{1,2\}} \left[\hat{\mu}_i(t) + \sqrt{\frac{8 \ln(1/\delta_t)}{T_i(t-1)}} \right]$$

Recall that Elegant UCB gives the following regret guarantee:

$$\overline{R_T} \leq \frac{2 \cdot 8 \ln(1/\delta_t)}{\Delta^2} + \sum_{t=1}^{T} \delta_t$$

3 Explore-then-exploit

The Algorithm for Explore-then-exploit is as follows:

1. Explore arm 1 in odd rounds and arm 2 in even rounds (for T_0 rounds each)
2. Always pick the best arm: $i \in \arg\max \hat{\mu}_i(T_0)$

Recall that $T_2(T)$ is the number of times that the inferior arm 2 is selected. If arm 1 comes out ahead after the first phase of the algorithm, arm 2 will never be selected again after the first $2T_0$ steps. However, if arm 2 wins, it will be chosen for the rest of the time. We can thus express this quantity as:

$$T_2(T) = \begin{cases} T_0 & \text{if } \hat{\mu}_1(T_0) \geq \hat{\mu}_2(T_0) \\ T - T_0 & \text{if } \hat{\mu}_1(T_0) < \hat{\mu}_2(T_0) \end{cases}$$

Taking expectations, we have:

$$\begin{aligned}\mathbb{E}[T_2(T)] &= \Pr[\hat{\mu}_1(T_0) \geq \hat{\mu}_2(T_0)](T_0) + \Pr[\hat{\mu}_1(T_0) < \hat{\mu}_2(T_0)](T - T_0) \\ &= (1 - \Pr[\hat{\mu}_1(T_0) < \hat{\mu}_2(T_0)])(T_0) + \Pr[\hat{\mu}_1(T_0) < \hat{\mu}_2(T_0)](T - T_0) \\ &= T_0 + \Pr[\hat{\mu}_1(T_0) < \hat{\mu}_2(T_0)](T - 2T_0)\end{aligned}$$

The event $\hat{\mu}_1(T_0) < \hat{\mu}_2(T_0)$ occurs only if one of the following events occurs:

1. $\hat{\mu}_1(T_0) < \mu_1 - \frac{\Delta}{2}$
2. $\hat{\mu}_2(T_0) > \mu_2 + \frac{\Delta}{2}$

We can upper bound the probability of the first arm performing worse than the second arm in the first $2T_0$ steps by applying the union bound as follows:

$$\begin{aligned}\Pr[\hat{\mu}_1(T_0) < \hat{\mu}_2(T_0)] &\leq \Pr[\hat{\mu}_1(T_0) < \mu_1 - \frac{\Delta}{2}] + \Pr[\hat{\mu}_2(T_0) > \mu_2 + \frac{\Delta}{2}] \\ &\leq 2e^{-\frac{\Delta^2 T_0}{8}}\end{aligned}$$

where the final inequality comes from the Hoeffding bound. Putting everything together gives us an upper bound on the expected number of times we pull the second arm:

$$\begin{aligned}\mathbb{E}[T_2(T)] &= T_0 + \Pr[\hat{\mu}_1(T_0) < \hat{\mu}_2(T_0)](T - 2T_0) \\ &\leq T_0 + 2e^{-\frac{\Delta^2 T_0}{8}}(T - 2T_0)\end{aligned}$$

Finally, we can plug in the definition of pseudo-regret:

$$\begin{aligned}\overline{R}_T &= \Delta\, \mathbb{E}[T_2(T)] \\ &\leq \Delta T_0 + 2\Delta e^{-\frac{\Delta^2 T_0}{8}}(T - 2T_0)\end{aligned}$$

In this expression, we can interpret the first term as the price of "exploration" and the second as the price of "exploitation." To see this more concretely, let's take a look at what happens for a couple different values of T_0.

Setting $T_0 = C$, a constant, corresponds to only exploring for a fixed number of time steps per arm regardless of the time horizon T. Intuitively, we expect our cost of exploration to be significantly lower than our cost of exploitation. Indeed, when we make the appropriate substitutions, we obtain the following expression for pseudo-regret:

$$C\Delta + 2\Delta e^{-\frac{\Delta^2 C}{8}}(T - 2C) = O(T)$$

The "exploration" part of the cost, $C\Delta$, turns into a constant while the "exploitation" part grows linearly in T. This is bad; in fact, asymptotically, this is just as bad as picking the worst arm every time.

Now, let's consider the opposite situation: $T_0 = \alpha T$ for some fixed $\alpha > 0$. Here, we let

the number of "exploration" steps grow linearly in the time horizon. After making the substitution, we have

$$\Delta(\alpha T) + 2\Delta e^{-\frac{\Delta^2 \alpha T}{8}}(T - 2\alpha T) = O(T)$$

The "exploration" part of the cost, $\alpha\Delta T$, is now linear in T while the "exploitation" part is negligible in comparison; the $e^{-\frac{\Delta^2 \alpha T}{8}}$ term amounts to dividing by an exponential in T, so it dominates the linear term it multiplies. Again, the pseudo-regret is linear in T.

Instead, "explore-then-exploit" selects T_0 to minimize pseudo-regret. We want to select a T_0 that grows asymptotically faster than a constant, but is also $o(T^a)$ for all $a > 0$. It turns out that the optimal value is $T_0 = \frac{c \ln(T)}{\Delta^2}$. Plugging this in, we get

$$\begin{aligned}
&\frac{c \ln(T)}{\Delta} + 2\Delta e^{-\frac{c \ln(T)}{8}}(T - \frac{c \ln(T)}{\Delta^2}) \\
\leq &\frac{c \ln(T)}{\Delta} + 2\Delta T^{-\frac{c}{8}} \cdot T \\
= &\frac{c \ln(T)}{\Delta} + 2\Delta T^{1-\frac{c}{8}}
\end{aligned}$$

Setting $c = 8$, for instance, makes the second term constant and gives a pseudo-regret guarantee of $O(\ln T)$ – much better than $O(T)$!

4 Lower Bound

Goal: Prove that there does not exist any approach that uses $\ll \ln T$ samples of worse arm regardless of $(\mu_1, \mu_2$

Definitions: E is a class of stochastic environments of $(\mu_1, \mu_2$ where $(\mu_1 \in [0,1], \mu_2 \in [0,1]$ this covers all stochastic 2-armed bandits.

A is a class of updates such that for every instance $(\mu_1, \mu_2 \in E : \mathbb{E}[T_{worse}(T)] = o(T^a) \forall a > 0$ i.e. "very sublinear"

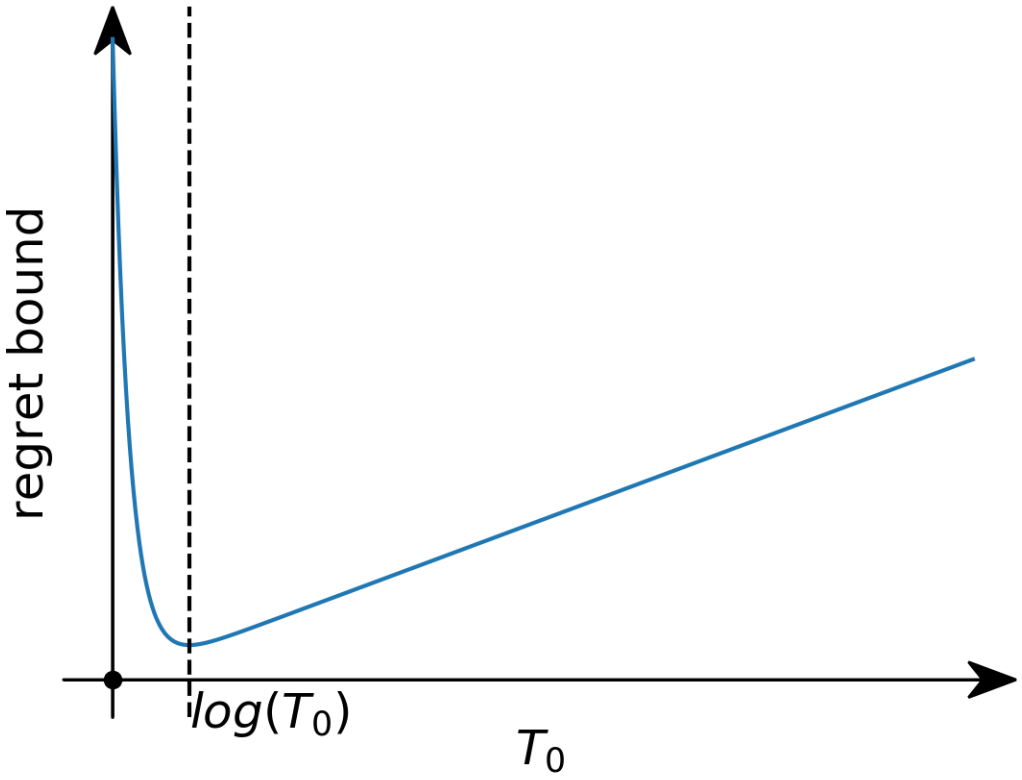
Lower bound (informal): For any env. $\in E$, any update $A \in A$ will pay

$$\overline{R_t} > \approx \frac{c\Delta \ln T}{D_{KL}(\mu_{worse} || \mu_{best}}$$

$$0 < p, q < 1 : D_{KL} = p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q} \geq \frac{1}{2}(p-q)^2$$

$$D_{KL}(p||q) \neq D_{KL}(q||p)$$

We do care about knowing the distributions of p and q; this lower bound does not hold when

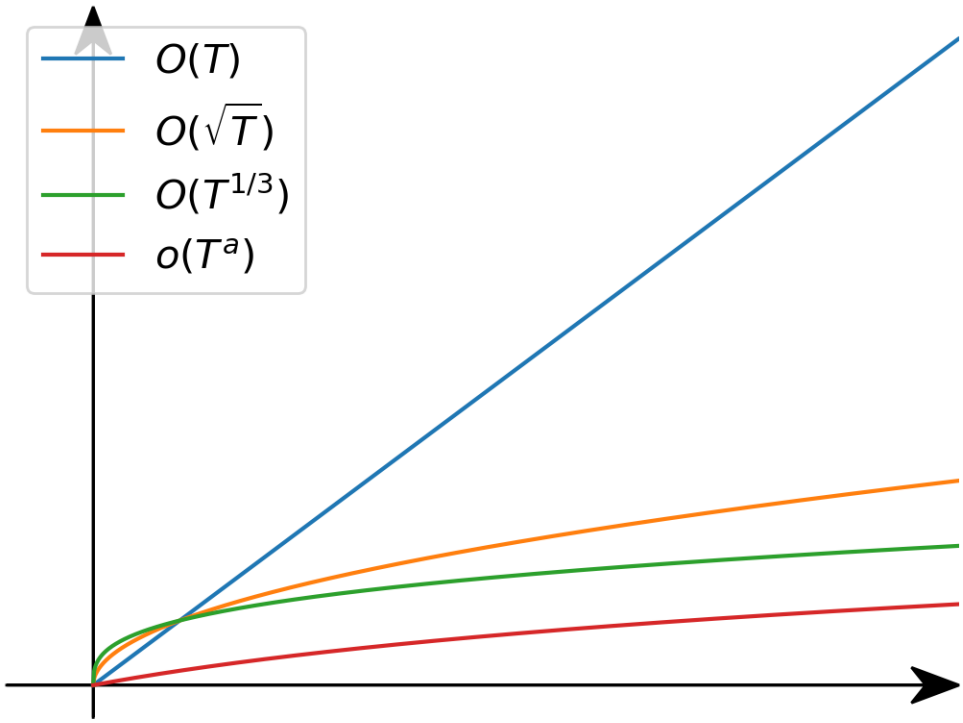


(a) Optimal $T_0 = O(\log T)$

we don't know the distribution, and in any case D_{KL} is asymmetric.

We're in universe A but could be in universe B, we don't know yet. If we sample a limited number of times then the confidence intervals for μ_{2A}, μ_{2B} will overlap.

Say $Z_1, Z_2 \cdots Z_n$ iid $Ber(q)$ so q is "true" distribution of our data, but if we inspect false environment by looking at sample mean $\frac{1}{n}\Sigma_{i=1}^{n} Z_i = p \neq q$ we want to bound our probability of getting "tricked".



(a) $o(T^a$ scales less than other common big-O

$$\begin{aligned}
\Pr_q[\frac{1}{n}\Sigma_{i=1}^{n} Z_i = np] &= \binom{n}{np} q^{np}(1-q)^{n(1-p)} \\
&= \binom{n}{np} e^{np\ln q + n(1-p)\ln(1-q)} \\
\frac{e^{nH(p)}}{2\pi\sqrt{n}} < \binom{n}{np} &= \frac{n!}{np!(n-np)!} < e^{nH(p)} \\
= e^{-np\ln p - n(1-p)\ln(1-p) + np\ln q + np(1-p)\ln(1-q)} & \\
&= e^{-nD_{kl}(p||q)}
\end{aligned}$$

Using Stirling's approximation to upper bound the binomial coefficient in the probability mass function of the Bernoulli random variable by the entropy of the data.

So, say we're in Universe B, the probability that we confuse ourselves by not sampling enough is lower-bounded by $e^{-nD_{KL}(p||q)}$ therefore we do need this minimum number of samples, which is not insignificant.

The specification of the Bernoulli distribution characterizing the reward of our arms is not important, what's important is that any time we have a finite number of samples we will have a blur around being able to pin down the parameter which characterizes the distribution of the reward of our arms.