So many decisions, so little time:
A Brief Introduction to Stochastic Multi-Armed Bandits

Christopher Kang
Click for Latest Version

November 7, 2021

1 Introduction & Key Ideas for Readers

This paper reviews the first chapter of Slivkin's book, *Introduction to Multi-Armed Bandits* [1]. I wanted to study multi-armed bandits (MABs) to understand how math tells us to make decisions. Currently, I am struggling to choose what to do after graduating - my hope is to study MABs and identify what strategies work and why. For readers, some key math ideas to keep in mind:

1. **Clean case / other case analysis** - our proofs will frequently use a clean case to simplify analysis. It's not always necessary to analyze all of the problems cases; sometimes, the worst case can be bounded and ignored via the law of total expectation.
2. **Weak bounds are okay**: the focus of this paper is upper bounding the regret incurred by the algorithms. It turns out that using weak upper bounds are often sufficient to achieve results.

2 Problem Context

2.1 Setup: So many choices, so little time!

In a multi-armed bandit, there is an agent with K different options ("arms") available to pull. Each arm a produces a specific reward from a distribution $\mathcal{D}_a$ (which, for convenience, will lie in $[0, 1]$). This agent then operates over T rounds, and typically $T \gg K$. Over each round, the agent only learns about the arm it pulled, and no other arms ("bandit" feedback). Thus, the key question driving our algorithms is - how do we trade off *exploration* of new arms with variable rewards vs *exploitation* of arms with predicted rewards?

Aside: The setup provided is kept simple intentionally. Other forms of multi-armed bandits include distributions which are dependent upon time, i.e. $\mathcal{D}_{a,t}$, or partial/full feedback systems where some information about the other arms is revealed.

2.2 Regret: What could have been

Because each arm could have rewards provided as a random variable, we first define the mean reward associated with a specific action a:

$$\mu(a) = \mathbb{E}[\mathcal{D}_a] \tag{1}$$

This yields a primary objective: minimizing expected regret, or $\mathbf{E}[R(T)]$. Regret is defined as:

$$R(T) = \mu(a^*) \cdot T - \sum_{t=1}^{T} \mu(a_t) \tag{2}$$

Where a^* is the "optimal" action, i.e.:

$$a^* \coloneqq \arg\max_a \mu(a) \tag{3}$$

So that $\mu(a^*)$ is the expected reward associated with the optimal action.

3 Strategies

3.1 Uniform exploration

The first strategy to try is just dedicating some period of time for exploration across all the arms, then spending the remaining time exploiting the highest value lever. Formally,

Uniform exploration over T total timesteps:

1. **Explore**: spend N trials on each arm
2. **Exploit**: spend the remaining $T - NK$ trials using the arm with the highest predicted mean rewards.

3.1.1 $K = 2$

Let's begin by identifying the expected regret when $K = 2$, i.e. there are two arms. Then, we can analyze the performance under two cases: the *clean case*, where our estimates of the rewards $\overline{\mu}(a)$ are close to the true mean $\mu(a)$, and the complement of the clean case.

Lemma 3.1. *Suppose we have some confidence interval $r(a) = \sqrt{\frac{2\log T}{N}}$. We define the clean event as:*

$$|\overline{\mu}(a) - \mu(a)| \leq r(a) \tag{4}$$

This occurs with probability exceeding $1 - \frac{2}{T^4}$.

Proof. Note that, because $\overline{\mu}(a)$ is the average of N independent samples from $[0, 1]$, we can simply apply Hoeffding's inequality:

$$\mathbb{P}[|\overline{\mu}(a) - \mu(a)| \geq r(a)] \leq 2\exp\Big(-\frac{2N^2 r(a)^2}{N}\Big) = 2\exp\Big(-2N\frac{2\log T}{N}\Big) = \frac{2}{T^4} \tag{5}$$

So the clean case then occurs with probability:

$$\mathbb{P}[|\overline{\mu}(a) - \mu(a)| \leq r(a)] \geq 1 - \frac{2}{T^4} \tag{6}$$

As desired. □

We will now estimate the regret $R(T)$ incurred in the clean case.

Lemma 3.2. *Assuming the multi-armed bandit is in the clean case, the regret will have an upper bound of:*

$$R(T) \leq N + \mathcal{O}\Big(T\sqrt{\frac{\log T}{N}}\Big) \tag{7}$$

Proof. There are two possibilities: either we select the optimal arm a^* after exploration or we select the suboptimal arm. Let's upper bound the regret incurred by selecting the suboptimal arm in the clean case. This would require that:

$$\mu(a) + r(a) \geq \overline{\mu}(a) > \overline{\mu}(a^*) \geq \mu(a^*) - r(a^*) \tag{8}$$

So, we can rearrange this quantity to find:

$$\mu(a^*) - \mu(a) \leq r(a) + r(a^*) \in \mathcal{O}\Big(\sqrt{\frac{\log T}{N}}\Big) \tag{9}$$

Note that this gives us a direct quantity for the regret incurred by selecting a instead of a^*. Thus, in this case, the regret scales in:

$$(N + T - 2N) \cdot (\mu(a^*) - \mu(a)) \leq N + \mathcal{O}\Big(T\sqrt{\frac{\log T}{N}}\Big) \tag{10}$$

Because $\mu(a^*) - \mu(a) \leq 1$ by the domain of the rewards. Note that this regret must be strictly less than the alternative, because in that case we only incur N incorrect arm pulls, whereas in this event we have incurred $N + T - 2N$ pulls. Thus, we have produced an upper bound, as desired. □

Now, we may complete our analysis. Specifically:

$$\mathbb{E}[R(T)] \leq \mathbb{E}[R(T)|\text{clean event}]\mathbb{P}[\text{clean event}] + \mathbb{E}[R(T)|\text{bad event}]\mathbb{P}[\text{bad event}] \tag{11}$$

$$\leq \Big(N + \mathcal{O}\Big(T\sqrt{\frac{\log T}{N}}\Big)\Big) + T \cdot \mathcal{O}(T^{-4}) \tag{12}$$

$$\leq N + \mathcal{O}\Big(T\sqrt{\frac{\log T}{N}}\Big) + \mathcal{O}(T^{-3}) \tag{13}$$

So, the contribution from the bad case is eclipsed by the contributions from the clean case. Note that we can seek to optimize that expression with ideal choice of N. By differentiation, we select $N = T^{2/3}(\log T)^{1/3}$:

$$\mathbb{E}[R(T)] \leq T^{2/3}(\log T)^{1/3} + \mathcal{O}\Big(T\frac{(\log T)^{1/3}}{T^{1/3}}\Big) \tag{14}$$

$$\in \mathcal{O}(T^{2/3}(\log T)^{1/3}) \tag{15}$$

3.1.2 Generalization

We can generalize the result identified in Equation (15) for $K > 2$:

Theorem 3.1 (Performance of the Uniform Exploration Algorithm)**.** *The expected regret scales in:*

$$\mathbb{E}[R(T)] \leq T^{2/3} \cdot \mathcal{O}(K \log T)^{1/3} \tag{16}$$

Proof. The generalized analysis builds on the 2 arm case. Note that we can compute the likelihood of the clean case. Recall that any arm could fail to be clean with probability $\leq \frac{2}{T^4}$, so union bounding yields an upper bound of any arms being unclean with probability $\frac{2K}{T^4}$ with maximum regret of 1.

Now, let's assume the clean case holds for all arms, which occurs with probability $\geq 1 - \frac{2K}{T^4}$. Then, we can take the same error bound derived, meaning that the regret from the exploitation phase scales in:

$$(T - KN)\mathcal{O}\Big(\sqrt{\frac{2\log T}{N}}\Big) \leq \mathcal{O}\Big(T\sqrt{\frac{\log T}{N}}\Big) \tag{17}$$

Including the exploration regret of at most $K \cdot N$, we have that total regret is upper bounded by:

$$\mathbb{E}[R(T)] \leq K \cdot N + \mathcal{O}\Big(T\sqrt{\frac{\log T}{N}}\Big) + \frac{2K}{T^4} \tag{18}$$

Selecting $N = (T/K)^{2/3} \cdot \mathcal{O}(\log T)^{1/3}$ minimizes this quantity:

$$\mathbb{E}[R(T)] \leq T^{2/3}K^{1/3}\mathcal{O}(\log T)^{1/3} + \mathcal{O}\Big(T\sqrt{(\log T)^{2/3}(T/K)^{-2/3)}}\Big) = \mathcal{O}(T^{2/3}(K\log T)^{1/3}) \tag{19}$$

□

How do we interpret this performance?: The Uniform Exploration algorithm has really poor scaling - first, this algorithm assumes that we have a fixed T. If we want to stop the algorithm before the exploitation phase, the average regret will be much higher because we've spent the entirety of the exploration phase accumulating regret. Second, the dependence on $T^{2/3}$ is subpar. The other algorithms we discuss will both be able to have arbitrary dependence on $t \leq T$ and better T, t scaling.

3.2 ϵ Greedy

The uniform exploration approach is okay, but what if we want to more evenly distribute the regret incurred by exploration? The ϵ greedy algorithm tackles this challenge by randomizing exploration.

ϵ **Greedy**: For each round t,

1. **Decide**: flip a coin with success probability ϵ_t
2. **Explore**: If success, choose an arm uniformly at random
3. **Exploit**: Otherwise, select the arm with the highest mean reward so far

In particular, we can set the dependence to be $\epsilon_t \sim t^{-1/3}$, meaning that as time goes on and we are more certain we have the best arm, exploration time is reduced.

3.2.1 Clean Event

To analyze this algorithm, we need to first create a new clean event ξ where:

$$\xi = \{\forall a, t : |\overline{\mu}(a) - \mu(a)| \leq r_t(a)\} \tag{20}$$

With confidence interval:

$$r_t(a) = \sqrt{\frac{2 \log T}{n_t(a)}} \tag{21}$$

Where $n_t(a)$ is the number of random samples performed on a. We'll need to clarify the argument before we can proceed, because $n_t(a)$ is also a random variable that's dependent upon previous interations. So, imagine that there is a "reward tape" for each arm, i.e. a $1 \times T$ tape, where each cell is a reward sampled from $\mathcal{D}_a$. We will say that the ith cell is the reward from the ith pull of that arm:

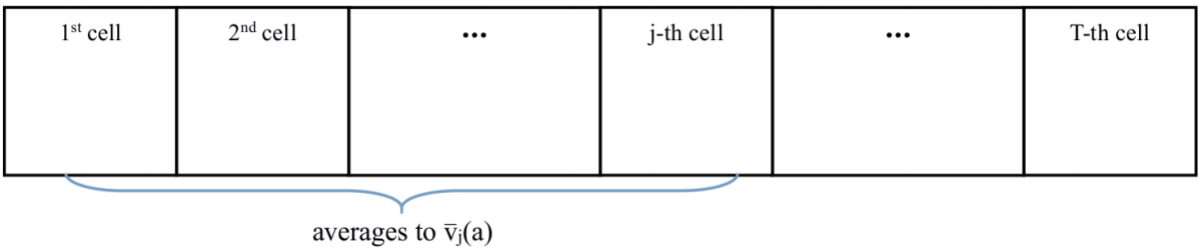

Figure 1: Book figure on reward tape visualization

We'll define $\overline{v}_j(a)$ as the average of the the jth pull of a. Then, by Hoeffding's inequality, we know that:

$$\forall j : \mathbb{P}[|\overline{v}_j(a) - \mu(a)| \geq r_t(a)] \leq 2 \exp\Big(-\frac{2j^2 r_t(a)^2}{j}\Big) = \frac{2}{T^4} \tag{22}$$

By union bounding over all of the arms (and assuming $K \leq T$):

$$\mathbb{P}[\forall a \forall j : |\overline{v}_j(a) - \mu(a)| \leq r_t(a)] \geq 1 - \frac{2KT}{T^4} \geq 1 - \frac{2}{T^2} \tag{23}$$

Which implies our clean event, as desired. Thus, we know that:

$$\mathbb{P}[\xi] \geq 1 - \frac{2}{T^2} \tag{24}$$

3.2.2 Analysis

Theorem 3.2. *The Epsilon Greedy algorithm has expected regret scaling in* $\mathcal{O}(t^{2/3} \cdot (K \log t)^{1/3}$.

Proof. Assuming the clean case, we aim to study the following quantity, where $\Delta(a_t) = \mu^* - \mu(a_t)$:

$$\mathbb{E}[R(t)] = \sum_{i=1}^{t} \mathbb{E}[\Delta(a_i)] \tag{25}$$

I.e. the expected per-round regret. Again, we can do a clean case / complement analysis, so we can focus on $\mathbb{E}[\Delta(a_i)|\xi]$. Again, exploration occurs with probability ϵ_i and can incur maximium regret of 1. However, the exploitation case is more complex - now, $r_t(a)$ $n_t(a)$ where $n_t(a)$ also a random variable. So, call N_t the total number of exploration rounds up to and including timestep t. Then, create event $\mathcal{B} \coloneqq N_t \geq \gamma \mathbb{E}[N_t]$ for some constant γ. Observe:

$$\mathbb{E}[N_t] = \sum_{k=1}^{t} \epsilon_k \geq t\epsilon_t \tag{26}$$

By the decreasing nature of ϵ_t. By the multiplicative Chernoff bound, we know that:

$$\mathbb{P}[\overline{\mathcal{B}}] \leq \exp(-(1-\gamma)^2 \mathbb{E}[N_t]/2) \leq \exp(-(1-\gamma)^2 t\epsilon_t/2) \tag{27}$$

We also need the event where the arm we select has a "fair" number of explores, i.e. $\mathcal{F}_j \coloneqq n_t(a_j) \geq \frac{N_t}{2K}$. Again by the multiplicative Chernoff bound:

$$\mathbb{P}[\overline{\mathcal{F}_j}] = \mathbb{P}[n_t(a_j) < (1-1/2)\frac{N_t}{K}] \leq \exp(-N_t/8K) \implies \mathbb{P}[\overline{F_j}|\mathcal{B}] \leq \exp(-\gamma t\epsilon_t/8K) \tag{28}$$

Now, we can analyze regret assuming $\mathcal{F}, \mathcal{B}$ hold for the exploitation phase. Recognize that:

$$\mathbb{E}[\mu(a^*) - \mu(a)|\mathcal{F} \wedge \mathcal{B}] \leq \mathbb{E}[r_t(a)|\mathcal{F} \wedge \mathcal{B}] + \mathbb{E}[r_t(a^*)|\mathcal{F} \wedge \mathcal{B}] \tag{29}$$

$$\leq 2\mathbb{E}\left[\mathcal{O}\Big(\sqrt{\frac{\log t}{n_t(a)}}\Big)|\mathcal{F} \wedge \mathcal{B}\right] \tag{30}$$

$$\leq \mathcal{O}\Big(\sqrt{\frac{K \log t}{\gamma t \epsilon_t}}\Big) \tag{31}$$

Segmenting over all the different cases we've analyzed (and noticing that $\overline{\xi}$ again drops out):

$$\mathbb{E}[\Delta(a_t)] = 1 \cdot \mathbb{P}[\text{Explored}] + \mathbb{E}[\Delta(a_t)|\mathcal{F} \wedge \mathcal{B}]\mathbb{P}[\mathcal{F} \wedge \mathcal{B}] + 1 \cdot \mathbb{P}[\overline{\mathcal{F}}|\mathcal{B}]\mathbb{P}[\mathcal{B}] + 1 \cdot \mathbb{P}[\overline{B}] \tag{32}$$

$$\leq \epsilon_t + \mathcal{O}\Big(\sqrt{\frac{K \log t}{\gamma t \epsilon_t}}\Big) + \exp(-\gamma t\epsilon_t/8K) + \exp(-(1-\gamma)^2 t\epsilon_t/2) \tag{33}$$

$$\leq \epsilon_t + \mathcal{O}\Big(\sqrt{\frac{K \log t}{\gamma t \epsilon_t}}\Big) + \mathcal{O}(\exp(-(1-\gamma)^2 t\epsilon_t/2)) \tag{34}$$

We are given $\epsilon_t = t^{-1/3} \cdot (K \log t)^{1/3}$ and $\gamma \in \mathcal{O}(1)$:

$$\leq \mathcal{O}(t^{-1/3} \cdot (K \log t)^{1/3}) + \mathcal{O}\Big(\sqrt{\frac{K \log t}{t^{2/3}(K \log t)^{1/3}}}\Big) + \mathcal{O}(\exp(-(1-\gamma)^2 t^{2/3}(K \log t)^{1/3}/2)) \tag{35}$$

$$\leq \mathcal{O}(t^{-1/3} \cdot (K \log t)^{1/3}) + \mathcal{O}\Big(\Big(\frac{K \log t}{t}\Big)^{1/3}\Big) + \mathcal{O}\Big(\frac{1}{t}\Big) \tag{36}$$

$$\leq \mathcal{O}(t^{-1/3} \cdot (K \log t)^{1/3}) \tag{37}$$

Each of the t rounds is asymptotically comparable to this one, so the regret scales in $\mathcal{O}(t^{2/3} \cdot (K \log t)^{1/2})$. (*Aside*: I am less confident about this proof - this was an exercise in the book and wasn't actually shown!) □

3.3 Adaptive Strategy: Successive Elimination

A key issue with both of the uniform exploration strategies is that arms suspected to be suboptimal are still explored. Thus, a simple strategy is to stop wasting resources on arms that we suspect to have little mean value (successive elimination).

3.3.1 Key Idea

The key idea is to use confidence intervals to eliminate arms with low mean value with high probability. If we know at any point that:

$$\overline{\mu}(a) - r_t(a) \geq \overline{\mu}(a') + r_t(a') \tag{38}$$

We can eliminate a' as being unlikely to have $\mu(a') > \mu(a)$ (seen below).

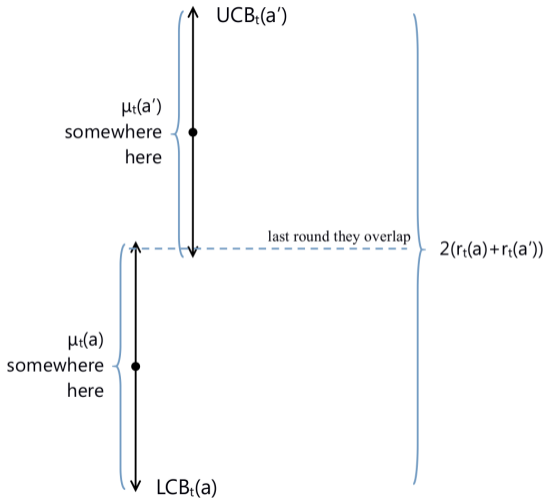

Figure 2: Book figure on confidence interval logic

3.3.2 Analysis over two arms

For simplicity, we again start with the case where $K = 2$:

Successive Elimination, $K = 2$:

1. **Explore**: Alternate testing each arm
2. **Exploit**: When one arm can be totally eliminated, only select the optimal arm forever

Theorem 3.3. *The successive elimination algorithm when $K = 2$ has expected error scaling:*

$$\mathbb{E}[R(t)] \leq \mathcal{O}(\sqrt{t \log T}) \tag{39}$$

Proof. Again, we will reference the clean event discussed in the ϵ greedy approach:

$$\xi = \{\forall a, t : |\overline{\mu}(a) - \mu(a)| \leq r_t(a)\} \tag{40}$$

In particular, note that the clean event necessitates that we select a^*, because otherwise we could not disqualify a. So, we simply need to estimate the regret that is required in order to identify a^*. Let's call t the last exploration round. By the above figure, we can see that:

$$|\mu(a^*) - \mu(a)| \leq 2(r_t(a^*) + r_t(a)) \leq 2\Big(\sqrt{\frac{2 \log T}{n_t(a^*)}} + \sqrt{\frac{2 \log T}{n_t(a)}}\Big) \tag{41}$$

Note that $n_t(a^*) = n_t(a) \approx t/2$ because exploration was evenly distributed; so, we know that:

$$|\mu(a^*) - \mu(a)| \leq 2\Big(\sqrt{\frac{2 \log T}{\lfloor t/2 \rfloor}} + \sqrt{\frac{2 \log T}{\lfloor t/2 \rfloor}}\Big) \in \mathcal{O}\Big(\sqrt{\frac{\log T}{t}}\Big) \tag{42}$$

So the total regret accumulated in the exploration phase is:

$$R(t) \leq |\mu(a^*) - \mu(a)| \cdot \frac{t}{2} \leq \mathcal{O}(\sqrt{t \log T}) \tag{43}$$

Including the bad case:

$$\begin{aligned} \mathbb{E}[R(t)] &= \mathbb{E}[R(t)|\xi]\mathbb{P}[\xi] + \mathbb{E}[R(t)|\overline{\xi}]\mathbb{P}[\overline{\xi}] && (44)\\ &\leq \mathcal{O}(\sqrt{t \log T}) + t \cdot \mathcal{O}(T^{-2}) && (45)\\ &\leq \mathcal{O}(\sqrt{t \log T}) && (46) \end{aligned}$$

As desired.

□

3.3.3 Generalizing with $K > 2$

This algorithm has an easy generalization to $K > 2$:

Successive Elimination: Start by marking all arms as "active." Then, every "phase,"

1. **Explore/Exploit**: Try all active arms
2. **Adapt**: Check all arms; if for any a there $\exists a' : \texttt{UCB}(a) < \texttt{LCB}(a')$, deactivate arm a.

Theorem 3.4. *Successive Elimination has a regret upper bounded by:*

$$\mathbb{E}[R(t)] \leq \mathcal{O}(\sqrt{Kt \log T}) \tag{47}$$

For all rounds $t \leq T$.

Proof. Again, we use a clean/complement type analysis. We use the same clean event ξ, which still occurs with probability exceeding $1 - \frac{2}{T^2}$. To bound the regret, we'll add the regret over the individual arms a that aren't the optimal a^*. So, consider the last round ℓ when a nonoptimal arm a is pulled. The regret is:

$$\Delta(a) \coloneqq \mu(a^*) - \mu(a) \leq 2(r_\ell(a^*) + r_\ell(a)) \tag{48}$$

By the UCB/LCB figure. Note that we've alternated all arms, so we know that $n_\ell(a), n_\ell(a^*)$ can differ by at most one. Furthermore, a is not played after ℓ, so $r_\ell(a) = r_T(a)$ Thus,

$$2(r_\ell(a^*) + r_\ell(a)) \in \mathcal{O}(r_\ell(a)) = \mathcal{O}(r_T(a)) \tag{49}$$

Thus, for any nonoptimal arm a (i.e. $\mu(a) < \mu(a^*)$):

$$\Delta(a) \leq \mathcal{O}(r_T(a)) = \mathcal{O}\Big(\sqrt{\frac{\log T}{n_T(a)}}\Big) \leq \mathcal{O}\Big(\sqrt{\frac{\log T}{n_t(a)}}\Big) \tag{50}$$

Because $n_T(a) \geq n_t(a)$ whenever $t \leq T$. Now, consider the set of all nonoptimal arms to be $\mathcal{A}^+$ and the set of all arms to be $\mathcal{A}$; the total regret incurred would be:

$$R(t) = \sum_{a \in \mathcal{A}^+} n_t(a)\Delta(a) \leq \sum_{a \in \mathcal{A}^+} \mathcal{O}(\sqrt{n_t(a)\log T}) \leq \mathcal{O}(\sqrt{\log T}) \sum_{a \in \mathcal{A}} \sqrt{n_t(a)} \tag{51}$$

Recall that $\sum_{a \in \mathcal{A}} n_t(a) = t$, because we've pulled a total of t arms in t rounds. Furthermore, $f(x) = \sqrt{x}$ is concave, so we can apply Jensen's inequality to upper bound this quantity (reference PS3 Q2 if you need a refresher). Noting that $|\mathcal{A}| = K$, we have:

$$\sum_{a \in \mathcal{A}} \frac{1}{K} \sqrt{n_t(a)} \leq \sqrt{\sum_{a \in \mathcal{A}} \frac{n_t(a)}{K}} = \sqrt{\frac{t}{K}} \tag{52}$$

So,

$$R(t) \leq \mathcal{O}(K\sqrt{\log T}) \sum_{a \in \mathcal{A}} \frac{\sqrt{n_t(a)}}{K} \leq \mathcal{O}(\sqrt{Kt \log T}) \tag{53}$$

We are almost done - the clean case now has bounded regret. Again, we can use a weak bound on the complement:

$$\mathbb{E}[R(t)] = \mathbb{E}[R(t)|\xi]\mathbb{P}[\xi] + \mathbb{E}[R(t)|\bar{\xi}]|\mathbb{P}[\bar{\xi}] \tag{54}$$

$$\leq \mathcal{O}(\sqrt{Kt \log T}) + t \cdot \frac{2}{T^2} \tag{55}$$

$$\in \mathcal{O}(\sqrt{Kt \log T}) \tag{56}$$

As desired. □

How do we interpret this performance?: UCB exhibits an immediate improvement with respect to t, T at a minor cost to K scaling. This is good news, because typically $t, T \gg K$, so our regret scaling is better than previous.

So, we have demonstrated three different algorithms - two which are non-adaptive and one which is adaptive - and analyzed their expected regret. This is just scratching the surface, but Successive Eliminatino demonstrates the power of adaptively adjusting to information and using clean cases to analyze problems.

4 Philosophy of Decision Making

4.1 Lessons

The algorithms we've discussed have a few key insights:

1. **Edge cases can be ignored**: Even though $\bar{\xi}$ can incur huge amounts of regret (we bounded it by t over t timesteps), the likelihood is so low that it doesn't make sense to actually plan your strategy around this case.

2. **Exploration is costly...**: as seen in the Uniform Exploration algorithm, the cost associated with searching all arms uniformly is what drives up the cost significantly.

3. **...So explore adaptively**: when comparing between the algorithms discussed, it's clear that adaptive is the way to go - if you know an arm has worse performance vs another arm, with high probability, you should not spend any more time exploring it.

Those who are familiar with the optimal stopping problem will notice some interesting differences between the strategies we've discussed. The secretary problem also has an exploration-exploitation tradeoff, but because exploitation occurs in a single phase, the optimal stopping policy is deterministic.

4.2 Challenges to real life

For me, studying MABs was a way to explore optimal decision policies to apply in my own life. Unfortunately, there are a few shortfalls:

1. **Problem setup**: notice that the above strategies vs the optimal policy for the secretary policy are drastically different - is life more like a MAB or optimal stopping problem? In addition, the MAB we considered had a number of simplifications: static $\mathcal{D}_a$, rewards in $[0, 1]$, fixed T. What if later mistakes have higher regret? What if the $\mathcal{D}_a \sim \mathcal{D}_a(t)$ (the distributions are changing over time)?

2. **Edge cases suck**: even though $\overline{\xi}$ is extremely unlikely to occur, it still is incredibly regretful. For example, a 0.1% chance of being homeless may seem insignificant until you actually are, well, homeless.

3. **Psychology of changing decisions**: even decisions that aren't regretful have decreasing utility by a happiness "set point" (see the hedonic treadmill). Does the strategy change if the rewards incentivize diversity of arm selection?

Does this mean the above analysis lacks life value? Absolutely not - the key lessons above still apply. Live life; explore aggressively, but be selective and willing to move on. Just keep in mind that the complexity of life necessitates more complex analysis than can be encapsulated in the MAB framework.

For further reading, the book mentioned has more chapters on MABs, including discussions on lower bounds and analysis over more complicated scenarios. There also exists a body of work on MABs focused on applications - for example, Netflix has significant research on leveraging MABs for recommendation algorithms. There are even connections to reinforcement learning! We have only scratched the surface.

References

[1] Aleksandrs Slivkins. *Introduction to Multi-Armed Bandits*. 2019. arXiv: 1904.07272 [cs.LG].