Online Multi-Armed Bandit
6.854 Final Project

Uma Roy, Ashwath Thirumalai, Joe Zurier
umaroy@mit.edu, ashwath@mit.edu, jayzee@mit.edu

December 14, 2016

Abstract

We introduce a novel variant of the multi-armed bandit problem, in which bandits are streamed one at a time to the player, and at each point, the player can either choose to pull the current bandit or move on to the next bandit. Once a player has moved on from a bandit, they may never visit it again, which is a crucial difference between our problem and classic multi-armed bandit problems. In this online context, we suggest several strategies and investigate their expected performance relative to optimal performance.

1 Introduction

Multi-armed bandit problems have been studied extensively in the literature. The classical version of the problem, often regarded as the canonical example of the exploration vs. exploitation tradeoff, is formulated as follows. There are N “bandits", an unknown set of distributions $\{D_i : 1 \leq i \leq n\}$, and a maximum number of allowed pulls, K. Each bandit can be pulled any number of times, and when bandit i is pulled, it provides a payout $p \simeq D_i$. The problem is formulated as sequentially deciding which of the N bandits to pull at each of the K stages. For the rest of this paper, we fix N and K to mean exactly the number of bandits and the number of pulls respectively. The bandit problem has many variants that have been explored previously in the literature. In this paper, we introduce a novel variant of the bandit problem and derive results analyzing several different strategies for our problem. However, we first provide some background about the general multi-armed bandit problem and several classical variants to appropriately place our work in relevant historical context and demonstrate what methods we adapt.

1.1 Background on the Multi-Armed Bandit Problem

The most obvious application of the classical multi-armed bandit problem is in online advertising. When firms like Google write algorithms to select which ads to show to a given user, these algorithms must decide whether to show an entirely new ad, or to show an ad for which the user's history indicates a higher likelihood of clicking. These algorithms are making the tradeoff between exploration and exploitation, deciding between which ads (bandits) to show the user (pull), without perfect information on the probability of the user clicking a given ad. As such, a lot of the research into multi-armed bandit problems has been conducted by Google and Yahoo Research, for example see [Sco15].

Many variants of this problem that fundamentally encapsulate a similar exploration versus exploitation tradeoff have been proposed and studied. Often times, these problems have similar applications as the classical version. A survey detailing many of the variants can be found at [MT08]–we describe a few particularly common variants here. One such variant is to assume the distributions D_i are Bernoulli distributions with unknown probability p_i of being 1, and that the values of p_i are drawn uniformly at random from [0, 1]. Known as the binary bandit or Bernoulli bandit problem, this level of specificity is most relevant in online advertising where the payoffs are binary (user either clicks or doesn't click on ad). Since the bandit payoff distributions are parametrized by a single variable, this variation simplifies a lot of analyses and allows for more meaningful theoretical bounding. Another variant, known as the pure exploration setting, focuses on simply finding the best bandit (one with the highest expectation) at the end of the designated number of pulls, with no penalty for the exploration phase [BMS09]. Yet another variant, known as the contextual bandit problem, includes a context vector that is provided to the player at each stage, and the player uses these context vectors and previous payoffs to decide which bandit to pull. Other variants include introducing correlations between the distributions of certain bandits, letting each bandit be a Markov bandit whose state evolves whenever it is played according to known Markov state evolution probabilities, allowing infinitely many arms (so instead of picking an integer between 1 and K, the player picks a real number), or allowing the player to choose two bandits at each stage and telling the player which bandit had the higher payoff. In this paper, we will primarily focus on a variant of the standard Bernoulli bandits.

The most common metric of success for a multi-armed bandit algorithm is its *regret*, which is the absolute difference between the expected total payoff of the algorithm and of an all-knowing algorithm. It's easy to see the expected total payoff of an all-knowing algorithm is Kd^*, where $d^* = max_i \overline{D_i}$, the maximal mean of the N bandit distributions and K is the number of pulls. While the regret is the best metric when no information about the distribution is known, knowing information about the distribution beforehand allows one to simplify the metric. In the case of the binary bandit problem where all n distributions are Bernoulli distributions, the expected value of the regret is very similar to the expected total payoff, and so it is essentially equivalent to look at the expected total payoff over all possible probabilities p_i.

The most common methods for solving multi-armed bandit problems include the ϵ-greedy strategy, the ϵ-first strategy, the Bayesian strategy, and the upper confidence bound (UCB) strategy. Most of these methods are summarized and empirically compared in [KP14]–again we briefly highlight a few important methods that we drew inspiration from in our own problem. The $\epsilon - greedy$ strategy, at each iteration, with probability ϵ, chooses a bandit to pull uniformly at random, and with probability $1-\epsilon$, chooses the bandit with the best average payoff thus far. The ϵ-first strategy involves pure exploration for the first $K\epsilon$ stages, meaning the player chooses a bandit to pull uniformly at random, and for the remaining $K(1-\epsilon)$ stages, the player pulls the bandit with the best performance thus far. In both these strategies, ϵ is essentially the proportion of the stages devoted to exploration as opposed to exploitation. The Bayesian strategy involves maintaining prior distributions (priors) of the expected value of each bandit, with the initial prior being $\text{Beta}(1,1) = U_{[0,1]}$ for each bandit. At each stage, for each bandit $1 \leq i \leq n$, the prior distribution of bandit i is sampled, and the bandit with the largest sample mean is chosen and pulled. The priors (which always remain Beta distributions) are then updated accordingly. This strategy, following Bayesian logic, never permanently discards losers, but instead pulls them at decreasing rates as the algorithm becomes increasingly sure of the winner. The final strategy, the upper confidence bound strategy, essentially establishes an upper bound for each bandit such that the probability that the expected value of the bandit is greater than the upper bound is negligible. It uses these upper bounds, with some adjustment, to determine which bandit to pull at each stage, and uses the Chernoff bound to bound the number of times that suboptimal bandits are played.

1.2 Our Problem

In this paper, we introduce and investigate the following online variant of the multi-armed bandit problem. This variant, to the best of our knowledge, is a *novel* variant of the classical multi-armed bandit problem, and all of our results in this area are original. Take the multi-armed bandit problem with Bernoulli (binary) bandits (as described above), and add the restriction that the bandits are streamed to the player one at a time. At each stage, the player decides whether to pull the lever on the current bandit. If the player decides to pull the lever, the stream is not advanced, and the player receives a payout from the current bandit drawn from its underlying distribution. If the player decides to not pull the lever, the stream is advanced and the player is presented with the next bandit. Once the stream reaches the last bandit, it stays there regardless of the player's actions, until the player has used up all K pulls. A key property of this formulation is that once the player decides not to pull a given bandit, he or she can never return to the same bandit and pull it again–making this problem fundamentally different from most classical formulations of the multi-armed bandit. For the purposes of our paper, we refer to the bandits as "coins" and pulls as "flips" interchangeably, for we can think of each bandit as a biased coin and each pull of a bandit as a flip of a coin.

Motivation for this formulation of the multi-armed bandit problem is derived from the famous secretary problem. Recall that the secretary problem consists of an agent determining the payoff of a stream of potential secretaries, and the agent must, at some point during the stream of secretaries, decide to hire a given secretary, with the goal of maximizing payoff. Similar to the secretary problem, this formulation of the multi-armed bandit problem involves deciding whether to continue with a given option (pull the lever on the bandit again) or proceed to the next one, with no possibility of ever returning. A crucial difference between our problem and the classical secretary problem is that we make the (strong) assumption that the means of our Bernoulli bandits are uniformly distributed on $[0, 1]$, meaning that our stream of bandits is randomly chosen, instead of by an adversary. We make this assumption so that we can give tight theoretical bounds for various strategies, which would not be possible in the case of an adversary. Instead of analyzing the classic measure of “regret", which measures performance of a strategy vs. an all-knowing algorithm (that would have payout $K * M$, where M is the maximal mean of the bandits), we simply evaluate the expected value of our strategy per pull. If we maintain the convention that getting payout 1 from a bandit is the same as flipping heads on the bandit (if we think of the bandits as biased coins), we see that the expected value of our strategy per pull is simply measuring how often we expect to flip heads given our strategy. We compare the expected value of our strategy per pull with 1–the absolute best number of heads one could hope to see per flip. One might note that if there are N bandits with means distributed uniformly on $[0, 1]$, the expected value for the maximum mean of the bandits would be $\frac{N}{N+1}$, so the optimal expected value per flip should be at most $\frac{N}{N+1} < 1$, but to simplify our measure of success, we always compare our expected value per pull with 1 as a baseline and note that we could easily compare it to $\frac{N}{N+1}$ if we chose to do so.

Many of the methods used in this paper are derived from solutions to either the multi-armed bandit problem, which we outlined in the previous section, or the secretary problem. While we analyze them for binary bandits, we are optimistic that similar methods/analyses should carry over for more general bandit distributions. Throughout this paper, we fix the following notation: N is the number of bandits (or coins) we are considering and K is the number of pulls (or flips) available.

We see right away that this problem has a number of properties that make it nice to analyze. Perhaps the most important is this: The entire current state can be specified knowing only the number of total lever pulls remaining, the number of bandits remaining, and the history of the current bandit. This makes it possible to write down recursive expressions for the optimal expected value.

2 Fixed Payout Bandits

Instead of immediately considering Bernoulli bandits, we will first consider **fixed payout bandits**, where each bandit i has some fixed payout μ_i for $\mu_i \in [0, 1]$. Like the Bernoulli bandits, we will assume that these fixed payouts are uniformly distributed on $[0, 1]$ so that we can provide theoretical analysis for the expected value of our strategies. Note that for the fixed payout bandit case, once we have pulled a bandit once, we have complete information about its payout distribution, unlike the Bernoulli bandit case, where uncertainty remains.

At first glance, one might think that the fixed payout bandit is equivalent to the rank optimization version of the secretary problem with N secretaries. However, this is false for two subtle reasons. Firstly, in the secretary problem we are only concerned with the relative rank of the person that we select. However here if the second highest mean bandit has a mean substantially lower than the maximal mean of the bandits, then our evaluation of our strategy penalizes this far more than the case where the second highest mean bandit has a mean almost equal to the maximal mean of the bandits. In the secretary problem, both of these strategies would be considered to give equal outcomes, i.e. it lacks a measure of distance, whereas our means inherently have a distance metric associated to them. The second reason is that any strategy we have incurs a penalty for any exploration that we do–i.e. if there are 2 bandits both with the maximal mean and if the first bandit we evaluate has the highest mean and we choose to skip over it but eventually find the second bandit with the maximal mean, and stay with it, this is worse than simply staying with the first bandit, since in the intermediate time we have been pulling on suboptimal bandits. However for the sake of the secretary problem, both of these situations would give the same result.

Given these subtle differences between the secretary problem and our problem, we now investigate the maximal expected payout solution to this problem. We do so by letting $f(K, N)$ be the expected payout of an optimal strategy for K pulls and N bandits. We derive a recursion for $f(K, N)$ and derive asymptotics for $f(K, N)$. Note that in deriving the recursion and asymptotics for $f(K, N)$, we never assume that the player is playing with a specific strategy–rather we only assume that the player is playing optimally. Thus our analysis of $f(K, N)$ doesn't provide an explicit strategy to use in this case, but rather provides an upper bound for what the expected value per flip of any strategy is. First we start with a lemma:

Lemma 2.1. *The payout of an optimal strategy satisfies the constraint that $\frac{f(K,N)}{K}$ is increasing (in K) for fixed N.*

Proof. We see that $\frac{f(K,N)}{K}$ represents the average payout per lever pull, given that the player has K lever pulls in total. Notice that at the last lever pull, the player has more information than at any other stage of the game, and so the expected payout of the last lever pull (which is purely exploitation) must be at least the average payout over the whole game. Therefore, the expected payout of the last lever pull is at least $\frac{f(K,N)}{K}$. Then, say the player instead has an extra lever pull, for a total of $K+1$ lever pulls. If the player simply executes the strategy for K, and then pulls the last lever once more, the player has increased the average payout over the whole game, and since the optimal strategy must be at least as good, this goes to show that the optimal average payout increases as K increases to $K+1$. Therefore, $\frac{f(K,N)}{K}$ is increasing over K with fixed N. □

Theorem 2.2. *Let p be the payout of the first bandit. Then, $f(K, N) = p + max((K-1)p, f(K-1, N-1))$.*

Proof. Since we don't have any prior information about specific bandits, it never makes sense to skip over a bandit without trying it, as the first bandit is functionally identical to all the other bandits initially. Therefore, the player will pull the first bandit, generating a payoff of p. Knowing the payoff p, if $p < \frac{f(K-1,N-1)}{K-1}$, the optimal strategy is to skip to the next bandit and recursively perform the algorithm for $f(K-1, N-1)$, because this would lead to a greater payout. If $p > \frac{f(K-1,N-1)}{K-1}$, then it makes sense to pull this bandit again. Then, we also know that $p > \frac{f(K-1,N-1)}{K-1} > \frac{f(K-2,N-1)}{K-2}$ by the above theorem, so it makes sense to pull the lever again. Continuing this argument, it makes sense to use all of the remaining lever pulls on this bandit. So therefore, we have that if $p < \frac{f(K-1,N-1)}{K-1}$, then $f(K, N) = p + f(K-1, N-1)$, and else, $f(K, N) = p + (K-1)p$, which is equivalent to $f(K, N) = p + \max((K-1)p, f(K-1, N-1))$, as desired.

□

Corollary 2.2.1. *Since $p \simeq U_{[0,1]}$, $f(K, N) = \int_0^1 p + \max((K-1)p, f(K-1, N-1))\ dp$.*

Theorem 2.3. *$f(K, N) = \frac{f(K-1,N-1)^2}{2(K-1)} + \frac{K}{2}$.*

Proof. If we let $E = \frac{f(K-1,N-1)}{K-1}$, we have that $f(K, N) = \int_0^E p + (K-1)Edp + \int_E^1 Kpdp = \frac{E^2}{2} + (K-1)E^2 + \frac{K}{2} - \frac{KE^2}{2} = \frac{(K-1)E^2}{2} + \frac{K}{2}$. Therefore, we can say that $f(K, N) = \frac{f(K-1,N-1)^2}{2(K-1)} + \frac{K}{2}$. □

We now investigate the asymptotic behavior of $f(K, N)$. It makes most sense to consider the convergence of $\frac{f(K,N)}{K}$ for fixed N as we increase K, or our number of pulls. Note that here and throughout the rest of this paper, we make use of the following terminology (which constitutes a mild abuse of notation): If two sequences a_n and b_n converge to some fixed K as $n \to \infty$, we say that a_n is asymptotic to b_n if the ratio $\frac{K-a_n}{K-b_n}$ converges to 1 as $n \to \infty$.

Theorem 2.4. *$\frac{f(K,N)}{K}$ is asymptotic to β_N for $K \to \infty$, where $\beta_N = \frac{1}{2} + \frac{1}{2}\beta_{N-1}^2$ and $\beta_1 = \frac{1}{2}$.*

Proof. It is clear that this quantity is always less than or equal to 1, since it the average payout per flip, and we showed it is increasing in Lemma 2, so therefore it must converge. Let's say it is asymptotic to β_N for $K \to \infty$. Then, substituting this in the theorem above gives us $K\beta_N = \frac{K-1}{2}\beta_{N-1}^2 + \frac{K}{2}$. Dividing by K, we have that $\beta_N = \frac{1}{2}\beta_{N-1}^2 + \frac{1}{2}$ (assuming $K \simeq K-1$ since $K \to \infty$), as desired. Moreover, it's easy to see $f(K, 1)$ for any K is equal to $\frac{K}{2}$, since it is just the expected value of the payout of the one bandit, so $\beta_1 = \frac{1}{2}$, as desired. □

Another intuitive way to see this recursion is that if you have infinitely many pulls, you will settle on pulling the first bandit permanently if its payout is greater than β_{N-1} (the payout per pull you'd get if you continued to the next bandit since you have $N-1$ bandits left), and otherwise proceed to the second bandit, giving you an expected payout of β_{N-1}. In the case of pulling the first bandit permanently, the expected value of its payout, given it is greater than β_{N-1}, is $\frac{1+\beta_{N-1}}{2}$, so the total expected value is $\beta_N = (1-\beta_{N-1})\frac{1+\beta_{N-1}}{2} + \beta_{N-1}\beta_{N-1}$, which simplifies to $\beta_N = \frac{1+\beta_{N-1}^2}{2}$, as desired.

We computed the first few numbers of the sequence β_i, which gives $\frac{1}{2}, \frac{5}{8}, \frac{89}{128}, \frac{24305}{32768}, \ldots$. After plugging these numerators into [OEI09] (sequence A167424), we found that someone had derived the asymptotics for β_i. In particular, [Fin03] gives the asymptotic value of β as $\beta_n \sim 1 - \frac{2}{n+\log(n)+O(1)}$.

Also notice that in the above section, we have not specified a strategy that actually achieves the optimal expected value of $f(K,N)$ for any K, N. Since we have derived a recursive closed form for $f(K,N)$ and given heuristics for asymptotic values of the sequence, one could use those values to actually play in such a game. However the main purpose of this section was to establish the best possible upper bound for the expected value of a strategy in the full information case (i.e. where we know the distribution of the bandits' payoff exactly after 1 pull). We shall see that this information will come in handy as an upper bound for any strategy we try in the Bernoulli bandits case in the sections below.

3 Bernoulli Bandits for Large Numbers of Flips ($K = O(N^c)$ for $c > 6$)

We start off by noting that in the Bernoulli bandits case, each bandit now has a payoff distribution Ber(p) for some parameter $0 \le p \le 1$, which is in some sense the 'bias' of the coin (if we think of bandits as biased coins). Now given one flip, we do not know the distribution of the payout of the bandit perfectly, making this scenario much more complicated. We also note that for K large and K small, relative to N, the streaming Bernoulli bandits problem is fundamentally different. As we increase the number of flips, it becomes less costly to explore our current bandit. Therefore for K sufficiently large, we should be able to approach the optimal strategy for the full-knowledge fixed payout bandit scenario since we can just pull each bandit sufficiently many times to get a good enough approximation of its parameter p, and then pretend that the bandit is a bandit with fixed payout p, since this is true in expectation. Indeed as $K \to \infty$, we see that this problem simplifies to the fixed payout bandit scenario using the strategy of pulling each bandit sufficiently many times to establish certainty about its empirical mean and then reducing to the fixed payout bandit scenario.

Since in the previous section, we asymptotically bounded the expected value per flip of the optimal strategy on the fixed payout bandits as β_n, which grows asymptotically as $1 - \frac{2}{n}$, we know that this serves as an upper bound for the expected value per flip of an optimal strategy for the Bernoulli bandits (since in this scenario we have strictly less information). Indeed, we can asymptotically reach this upper bound given enough lever pulls and provide an explicit description of such a strategy that approaches the asymptotic limit. Consider the following strategy given $K \ge N^{6+\epsilon}$:

Definition 1 (K large Bernoulli bandit strategy)**.** *When we're on the i^{th} bandit $(1 \le i \le n)$, pull each lever a large number of times $(N^{4+\epsilon/2})$ and continue pulling it for the remainder of the game if the empirical mean of the bandit after this time is $\ge \frac{N-2-(i-1)}{N-(i-1)}$. Otherwise, move on to the next bandit.*

Theorem 3.1. *The strategy described above results in a payout asymptotic to $K(1-\frac{2}{N})$, or $1-\frac{2}{N}$ payout per flip.*

Proof. Let $P(i)$ denote the probability that we stick with bandit i, and $E(i)$ denote the expected payout from repeatedly playing bandit i given that we stick with it. We seek to bound the sum $\sum_{i=1}^{N} P(i)E(i)$ from below. Note that $P(i)E(i) \ge Q(i)F(i)$, where $Q(i)$ denotes the probability that we stick with bandit i and that it has an underlying payout greater than or equal to $\frac{N-2-(i-1)}{N-(i-1)} + N^{-\alpha}$, and $F(i)$ denotes the expected payout from repeatedly pulling bandit i given that its payout is at least this large. (Here $\alpha = 2 + \frac{\epsilon}{5}$; the reason for this value is so that $N \cdot N^{-\alpha} = o(N^{-1})$). We also observe that the total number of lever pulls used for discovery is $\le N(N^{4+\epsilon/2}) \le \frac{K}{N^{1+\epsilon/2}}$, so the number of lever pulls spent on repeatedly pulling bandit i is at least

$K(1-N^{-1-\epsilon/2})$. Hence $F(i) \geq K(1-N^{-1-\epsilon/2})\cdot\frac{1}{2}\left(\frac{N-2-(i-1)}{N-(i-1)}+1\right)$. (Note that we are averaging the endpoints of the expected range of probabilities since the means are uniformly distributed.)

It remains to bound $Q(i)$. Let $A(i)$ denote the probability that we move on from bandit i given that we get there, and let $B(i)$ equal the probability that we stay on bandit i and that it has underlying payout $\geq \frac{N-2-(i-1)}{N-(i-1)} + N^{-\alpha}$. Note that $Q(i) = B(i)\prod_{j=1}^{i-1} A(i)$, so it suffices to bound A and B separately using the Chernoff bound. Intuitively, we would expect $A(i) \approx \frac{N-2-(i-1)}{N-(i-1)}$ and $B(i) \approx 1 - \left(\frac{N-2-(i-1)}{N-(i-1)} + N^{-\alpha}\right)$. Indeed, using the Chernoff bound we can easily compute that $A(i) \geq \frac{N-2-(i-1)}{N-(i-1)} - N^{-\alpha}\exp(-C_1N^{\epsilon/10})$, where C_1 is some positive constant. Similarly, $B(i) \geq 1 - \left(\frac{N-2-(i-1)}{N-(i-1)} + N^{-\alpha}\right) - \exp(-C_2N^{\epsilon/10})$.

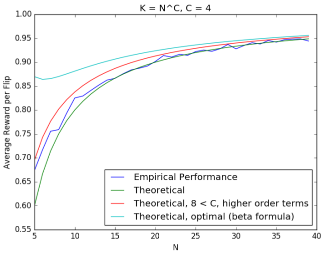
We therefore deduce that $Q(i) \geq (1-\frac{N-2-(i-1)}{N-(i-1)})\prod_{j=1}^{i-1}\frac{N-2-(j-1)}{N-(j-1)} - o(N^{-1})$. Given this product and the bound on $F(i)$, we can compute the desired asymptotic expression using Mathematica. Mathematica evaluates $(1-\frac{N-2-(i-1)}{N-(i-1)})\prod_{j=1}^{i-1}\frac{N-2-(j-1)}{N-(j-1)}$ to $\frac{2(N-i)}{N(N-1)}$; we can drop the $o(N^{-1})$ lower order term for the remainder of this analysis. Hence our expectation is given (asymptotically speaking) by

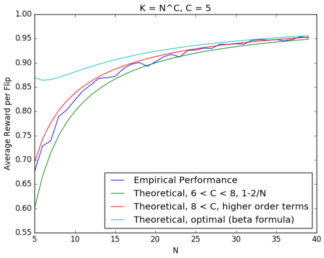
$$\sum_{i=1}^{N}\left(\frac{2(N-i)}{N(N-1)} - o(N^{-1})\right) K(1-o(N^{-1})\cdot\frac{1}{2}\left(\frac{N-2-(i-1)}{N-(i-1)}+1\right)$$

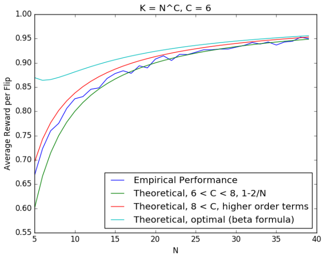
$$\sim K(1-\frac{2}{N})+o(\frac{1}{N})$$

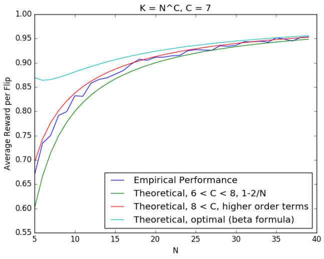
again making use of Mathematica to evaluate the sum and provide asymptotics. □

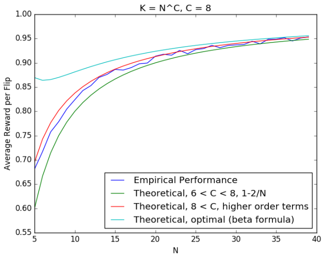
We make a couple of comments about the theorem and proof above. Firstly, note that for $C > 8+\epsilon$, we can recover the next lower order term in the asymptotic expression for β ($\beta_n \sim 1 - \frac{2}{n} + \frac{2\log(n)}{n^2} + O(\frac{1}{n^2})$ by following the above analysis with $N^{6+\epsilon/2}$ flips per bandit. Also, we see that the tail bounds derived from the Che7rnoff inequality in our proof result in the necessary condition that $C > 6$ in order for our result to hold. However, intuitively we do not need such a strong condition–empirically we see that $C \geq 4$ (adjusting the number of times we flip each level to be N^2 when $K = N^4$) suffices for our result to hold. Indeed, we wrote an empirical simulation of our strategy and tested it for fixed $N = 10$ and various $K = N^c$ for various c and plotted our empirical mean payout per flip. These graphs are below.

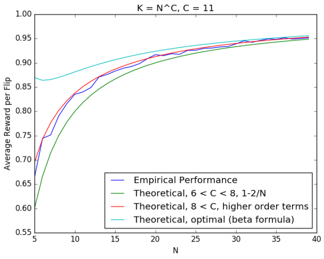












We make a few interesting observations about the graphs above. For $C \geq 8$, we see that the strategy converges to the theoretical bound with the lower order terms, as expected. For $8 > C \geq 6$, we see that the strategy converges to the theoretical bound without the lower order terms. Finally for $C = 4, 5$, these graphs show empirically that the strategy is still asymptotically converging to the theoretical bound for $C \geq 6$, although we could not tightly prove this. Although graphs for $C < 4$ are not shown, in the range $N \leq K \leq N^3$, this strategy does poorly so this is where we focus next. As we shall see, this scenario is a fundamentally different case.

4 Bernoulli Bandits for Small Numbers of Flips

In the previous section, we considered Bernoulli bandits in the case that our number of pulls, K was much greater than the number of bandits N. Intuitively, this allowed us to pull each bandit many times to establish a good sense of the Bernoulli parameter of the bandit, and evaluate whether it was better than the expected maximum of the remaining bandits. Since the number of pulls was much greater than the number of bandits, this "exploration" of establishing certainty about the bandits' distribution did not detract too much from our final payoff, allowing our strategy to be asymptotically optimal. The much more interesting case is when K is small relative to N.

In the case that our number of flips $K = N$, a similar strategy to the previous section does not work. If we spend too much time "exploring" then we cannot spend any time at all "exploiting" the good bandits that we find. The first such strategy we analyze is the following:

Definition 2 (Tails Strategy)**.** *When you first get a bandit, pull it until you see a tails. Once you see a tails, move on to the next bandit.*

This strategy is quite simple indeed–we simply play each bandit until we see an unfavorable outcome (a tails) and then move on to the next bandit. Note that this strategy only is feasible for K not much larger than N (say, around $\Theta(N\log(N))$), since for larger K, we have a constant probability of arriving to the last bandit, which will determine our payout, and since the expected value of the last bandit is $\frac{1}{2}$, our average payout will be dominated by a $\frac{1}{2}$ term, which is bad.

Theorem 4.1. *Set $K = N$. Then, the tails strategy has expected payout asymptotic to $N(1 - \frac{1}{\log N})$ as $N \to \infty$.*

Proof. We first argue that with F flips remaining and $\geq F$ bandits remaining, the expected average payout per flip is asymptotic to $(1 - \frac{1}{\log(F)})$. To see why, note that we may compute the average payout per bandit as follows: $\sum_{i=1}^{F} P(i)$, where $P(i)$ denotes the probability of getting at least i heads in a row. Clearly $P(i) = \int_0^1 p^i \mathrm{d}p = \frac{1}{i+1}$. Meanwhile, the average number of flips used per bandit is simply $1 + \sum_{i=1}^{F} P(i)$, since we always use exactly one more flip than we get heads (as the last flip is a tails and all preceding flips are heads). Hence the payout per flip in this case is equal to $\frac{H_F - 1}{H_F}$, where $H_F = \sum_{i=1}^{F} \frac{1}{i} \sim \log(F) + \gamma$. Therefore we see that the payout per flip is asymptotic to $1 - \frac{1}{\log(F)}$. Now consider the first $K(1 - \frac{1}{\log(K)^2})$ flips of the tails strategy. Since for these flips the total number of flips remaining is always at least $\frac{K}{\log(K)^2}$, we deduce that the expected payout per flip over this period is at least $(1 - \frac{1}{\log(K/\log(K)^2)})$. Hence the total payout over this period (and thus the total payout of the tails strategy for all flips, not just the first $K(1 - \frac{1}{\log(K)^2})$ is equal to $K(1 - \frac{1}{\log(K)^2})(1 - \frac{1}{\log(K) - 2\log(\log(K))})$, which as desired is asymptotic to $K(1 - \frac{1}{\log(K)})$.

To show that the tails strategy is no better than this, note that our average number of heads per flip over the entire course of the strategy is bounded above by the average number of heads per flip when we have all K flips remaining, which by the above is simply $(1 - \frac{1}{\log(K)})$. □

As a side note, we explore a recursive way to prove the above theorem that has some interesting analysis. This recursion gave good heuristic results so we simply state the intuition here for the interested reader. If we let $T(k)$ denote the expected payout of the tails strategy given k flips and $\geq k$ bandits. Note that once there are at least as many bandits as flips, as long as we don't skip any bandits this remains true for the remainder. We derive the following recursion for $T(k)$: First note that the possible sequences of heads and tails we may encounter on this bandit are $H^iT : 0 \leq i \leq k-1$ and H^k. The probability of this latter case is $\int_0^1 p^k \mathrm{d}p = \frac{1}{k+1}$; since our payout in this case is k, the contribution to the expectation is $\frac{k}{k+1}$. For the former case, the probability is given by $\int_0^1 p^i(1-p)\mathrm{d}p = \frac{1}{(i+1)(i+2)}$, and the payout in this case is $i + T(k - (i+1))$ (since we have i heads and have $k - (i+1)$ flips remaining). We deduce that

$$T(k) = \sum_{i=0}^{k-1} \frac{1}{(i+1)(i+2)}(i + T(k-i-1)) + \frac{k}{k+1}$$

.

Empirically, we see that $T(k)$ is bounded by $k - \frac{k}{\log(k)}$ (when we plot both quantities). If we do some heuristical analysis and plug in $T(k) = k - \frac{k}{\log(k)}$ into the recursive definition, we see that equality is obvious, however making this equality rigorous instead of manipulating limits a la Euler proved beyond our means. Thus we see that $\frac{T(k)}{k}$, which is our expected payout per flip for this strategy is asymptotic to $1 - \frac{1}{\log(k)}$ as $K \to \infty$, matching the theorem above.

Considering the naivete of such a simple strategy, it is remarkable that is is only $\frac{1}{\log N}$ off from the absolute optimal case, where each flip we do, we see a heads (a payout per flip ratio of 1). One might attempt to generalize this strategy by specifying a parameter α such that instead of moving on immediately when we see a tails, we instead move on when the empirical proportion of heads we have seen so far is less than the threshold α. This motivates the following definition of a strategy:

Definition 3 (α-tails strategy)**.** *Each time you pull a bandit, keep track of the number of heads H you have seen and the number of tails T you have seen. If $H + T > 0$ (i.e. you have pulled the bandit at least once) and $\frac{H}{H+T} \geq \alpha$, then pull the current bandit again. Otherwise, move on to the next bandit.*

Note that the tails strategy is equivalent to setting $\alpha = 1$, as we have that we move on when the number of heads H and the number of tails T satisfy $\frac{H}{H+T} < 1 \iff 0 < T$, i.e. we move on whenever we see a tail. By fixing a threshold α more intelligently than simply setting $\alpha = 1$, it is conceivable that we may recover a strategy with better asymptotic expected payoff per flip than $1 - \frac{1}{\log(n)}$. However, we prove the following theorem about such a class of strategies:

Theorem 4.2. *Fix $0 \leq \alpha \leq 1$ as the threshold for an α-tails strategy. Then the expected value per flip of such a strategy is upper bounded by $\frac{\alpha+2}{3}$.*

Proof. The intuition here is that our payout per flip from a given bandit is bounded above by α if we move on from that bandit, by definition. Therefore the only way to exceed α would be to remain at one bandit for the rest of the game. Here we would expect to get a payout of roughly $\frac{\alpha+1}{2}$ per flip, since the bandits that we would stay at for an extended period of time will have means distributed in the interval $[\alpha, 1]$. We can make this proof rigorous by providing an upper bound on the expected payout per flip of a bandit that we pull K times. This can be done by computing the component of the expected value of the bandit that comes from cases in which we pull the lever K times, then dividing by both K (to convert to a per flip value) and the probability that we pull the lever K times on our bandit. We claim that the first of these quantities (the expected value component) is given by the following expression:

$$\sum_{a=\lceil \alpha K \rceil}^{K} \left[\int_0^1 p^a (1-p)^{K-a} \binom{K}{a} \frac{a - (K-a)(\frac{\alpha}{1-\alpha})}{K} \mathrm{d}p \right] a$$

To see why this is true, we recall that the valid sequences of heads and tails are precisely those that solve the generalized ballot problem: the running total of heads must exceed the number of tails by a factor of $\frac{\alpha}{1-\alpha}$ at each step in order for the ratio of heads to flips to exceed α. We then multiply by the number of such sequences $\binom{K}{a}$ and the probability of each one occurring $p^a(1-p)^{K-a}$. Integrating over the possible p and multiplying by the payout a gives the desired expression. Note that the lower bound on the sum is due to the fact that we must clearly have at least αK heads in order to satisfy the ballot condition. We then divide this sum by the probability that this event occurs, which is quite clearly given by simply omitting the rightmost a above (i.e. summing over probabilities without weighting by payout). We can compute this ratio explicitly using Mathematica (which is able to evaluate these sums if we marginalize out the integral first); the result is an expression that goes asymptotically to $\frac{K(2+\alpha)}{3}$ as claimed.

□

If we set $\alpha < 1$ in this theorem, we simply get that the expected value per flip of the fixed α strategy is upper bounded by $\frac{\alpha+2}{3}$, which is some fixed constant < 1. Thus we see that setting any constant $\alpha < 1$ is inferior to setting $\alpha = 1$, or the tails strategy, since that has expected value per flip $1 - \frac{1}{\log(N)}$, which converges to 1 as $N \to \infty$. (Note that if we set $\alpha = 1$ in the above theorem, we only get that the expected value per flip of the tails strategy is upper bounded by $\frac{1+2}{3} = 1$, which gives us no non-trivial upper bound). Somewhat surprisingly, we have shown that $\alpha = 1$ is the best possible choice of α if we are using a fixed α–“keep it simple, stupid" really holds true in this case!

If a constant α does not lead to asymptotically good results, the natural next step is to make α a function of N. Note that for simplicity we still fix α once we have chosen N; i.e. we do not decrease our threshold during our play, even as we are left with fewer and fewer remaining bandits. With this greater flexibility, we can improve upon the $1 - \frac{1}{\log(N)}$ upper bound we get on the expected value of the tails strategy. The following theorem states this result:

Theorem 4.3. *Let $K \leq N$ and set $\alpha = 1 - \left(\frac{\log(N)}{N}\right)^{\frac{1}{3}}$. The resulting strategy has an expected value $\geq K(3\alpha - 2)$; i.e. our expected payout per flip is $\geq 1 - 3\left(\frac{\log(N)}{N}\right)^{\frac{1}{3}}$. Furthermore, we can take K to be as large as $N^{3-\epsilon}$ for $\epsilon > 3$ and achieve the same bound.*

Proof. Fix $\tau = \frac{\alpha}{1-\alpha}$. First, we note that by a previous computation, the probability that we flip the lever on a given bandit for the remainder of the game (supposing we have K lever pulls left) is given explicitly by the sum

$$\sum_{a=\lceil \alpha K \rceil}^{K} \left[\int_0^1 p^a (1-p)^{K-a} \binom{K}{a} \frac{a - \tau(K-a)}{K} \mathrm{d}p\right]$$

$$\geq \sum_{a=\alpha K}^{K} \frac{a - \tau(K-a)}{K(K+1)}$$

$$= \frac{1 + K(1-\alpha)}{2(1+K)}$$

which for K sufficiently large exceeds $\frac{1-\alpha}{3}$. We call a bandit *strong* if we end up pulling its lever for the rest of the game. Note that a strong bandit has a payout per flip at least as good as α.

We next observe that if we pull the lever on a given bandit a total of L times, then our payout must be at least $(L-1)\alpha$ (otherwise we would have moved on already). Therefore our payout per flip for this bandit is $\frac{(L-1)}{L}\alpha$, so if $L \geq \frac{1}{1-\alpha}$ we deduce a payout per flip of at least $\alpha^2 \geq 2\alpha - 1$. This is better than the bound we are trying to prove, so we do not need to worry about such bandits.

The probability that we do not find any strong bandits among the first $\log(N)^{\frac{2}{3}} N^{\frac{1}{3}}$ is given (due to independence) by $(1 - \frac{1-\alpha}{3})^{\log(N)^{\frac{2}{3}} N^{\frac{1}{3}}} \leq N^{\frac{-1}{3}}$. Supposing we do find a strong bandit in this range, the number of "wasted flips" (flips that could have possibly been exercised on a bandit with a net payout per flip smaller than $2\alpha - 1$) is bounded above by $\log(N)^{\frac{2}{3}} N^{\frac{1}{3}} \cdot \frac{1}{1-\alpha} = \log(N)^{\frac{1}{3}} N^{\frac{2}{3}}$. Putting everything together, we see that with probability $(1 - N^{-\frac{1}{3}})$ the payout per flip is some convex combination of the quantities $2\alpha - 1$ (from bandits that we pull more than $\frac{1}{1-\alpha}$ times, not including the strong bandit) and $\alpha \frac{K - \log(N)^{\frac{1}{3}} N^{\frac{2}{3}}}{K} = 2\alpha - 1$ (where the first factor comes from the payout per flip of the strong bandit, and the second factor accounts for wasted flips). Since we are multiplying $2\alpha - 1$ through by the $(1 - N^{-\frac{1}{3}})$ probability of our setup working correctly, we deduce a net expected value per flip of $\geq 3\alpha - 2$ (since $N^{-\frac{1}{3}}$ is dominated by $\left(\frac{\log(N)}{N}\right)^{\frac{1}{3}}$).

The claim about K larger relative to N is easy to deduce from the above proof, since we consider only cases in which we make use of the first $\log(N)^{\frac{2}{3}} N^{\frac{1}{3}}$ bandits. Also note that we make no attempt to optimize the constant factor 3 used in the above argument. □

5 A Look At the Optimal Strategy

In this section we state a recursive formula that we derive for optimal play that has proven useful and empirically plot this quantity for benchmarking purposes. This optimal payout serves as an upper bound for the payout of any strategy and is not derived from applying some specific strategy.

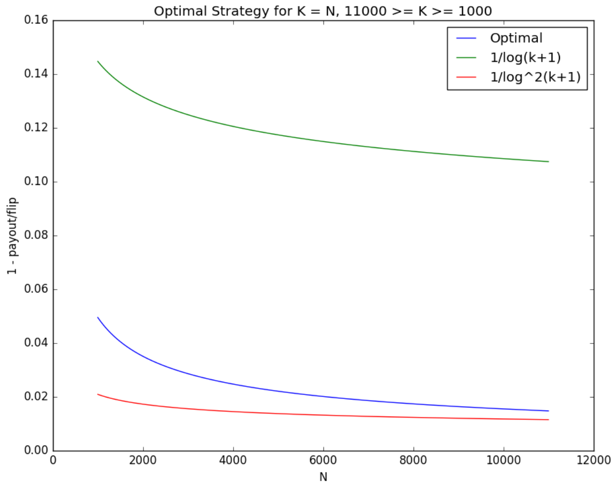
5.1 Recursive Expression for Optimal Play

We derive a recursive expression for optimal play with K flips, N bandits. Let $O(K, N, a, b)$ represent the maximal optimal payout if you have K flips left, N bandits left (after your current bandit) and on your current bandit you have flipped a heads and b tails. We see that the optimal payout for K flips and N bandits is simply $O(K, N, 0, 0)$, i.e. when you start you have not flipped anything on your first bandit and you have all remaining flips and N bandits to examine.

Our base cases are as follows: if $K = 0$, then $O(0, N, a, b) = 0$, as you have no flips left. If $N = 0$, this means you have no bandits left. i.e you are on the last bandit. Thus your payout is simply the expected payout of the last bandit (which is simply the mean of the $\beta(a + 1, b + 1)$ distribution since that is the distribution for the bandit given your observations) times the number of flips you have left, which means $O(K, 0, a, b) = K\frac{a+1}{a+b+2}$. Finally, if both N and K are non-zero, the probability of flipping heads is $\frac{a+1}{a+b+2}$, in which case our payout is $1 + O(K - 1, N, a + 1, b)$, since we get a 1 from the head we have just flipped, and our state advances to have exactly 1 less flip, and we stay on the same bandit. With probability $\frac{b+1}{a+b+2}$, we flip a tails, in which case our payout is the maximum of $O(K-1, N-1, 0, 0)$, which represents moving on to the next bandit, and $O(K - 1, N, a, +1)$ which represents staying on the current bandit and updating the observations for the bandit. Thus if $K, N \neq 0$, we have that $O(K, N, a, b) = \frac{a+1}{a+b+2}(1 + O(K - 1, N, a + 1, b)) + \frac{b+1}{a+b+2}\max(O(K - 1, N_1, 0, 0), O(K - 1, N, a, b + 1))$, since we are calculating the expected value of the payout.

By providing a recursive formula for the maximum possible payout for optimal play, we can compute $O(K, N, 0, 0)$ for any K, N of our choosing and use these computations as a benchmark for any strategy that we wish to test. Since for the case where K is large compared to N, we have a provably asymptotically optimal strategy, we mostly computed $O(K, N, 0, 0)$ for $K = N$ and increasing N to benchmark the tails-strategy's performance (which we computed via the recursive formula given after the proof of Theorem 4.1) against optimal payout. For many other strategies that we tried, we were able to derive similar recursive formula for their expected performance, which we plotted against the optimal payout to see which strategies were better.

Below is a graph showing the values of $1 - O(K = N, N, 0, 0)/N$ for N from 1 to $11,000$. Note that this quantity is worth plotting because $O(N, N, 0, 0)/N$ measures our payout per flip, and $1-$ payout per flip is measuring how far off we are from being perfect (i.e. flipping a head each time). We see that $O(N, N, 0, 0)$ decays extremely quickly towards 0. We attempted to find asymptotics for $O(N, N, 0, 0)$, but it was difficult to find any obvious fit, and given any fit it was almost impossible to analyze given the 4 variable nature of this recurrence. Finding asymptotics for this quantity would be quite interesting work for the future.



6 Conclusion and Future Work

In this paper, we considered a novel (to our knowledge) variation of the multi-armed bandit problem that combined the idea of the fundamental exploration-exploitation tradeoff present in bandit problems with the idea of streaming input from the classic secretary problem. We started with the problem where the bandits had fixed payout distributions, and in this case derived an asymptotic upper bound on the payout of the optimal strategy, which served as an upper bound for any strategy in the Bernoulli bandit case. For the Bernoulli bandit case for K large, we were able to provide an asymptotically optimal strategy that recovered the payout for the fixed point bandits for $K > N^c$ for $C \geq 6 + \epsilon$. For the Bernoulli bandit case for K small, on the order of $O(N^3)$ or less, we were able to describe a naive strategy that provably does well and has expected payoff per pull asymptotically equal to $1 - \frac{1}{\log(N)}$ as $N \to \infty$. Furthermore, we elaborated on this strategy to derive a strategy with asymptotic expected payoff per pull equal to $1 - 3\left(\frac{\log(N)}{N}\right)^{1/3}$, which was an improvement over our first naive strategy.

While we have provided provably good strategies for many cases, there is much room for further exploration. First off, in Section 5, we derive and compute a recursion for the optimal strategy for $K = O(N)$. Proving asymptotics for this recursion would be quite interesting, although extremely difficult because the recursion is in 4 variables. If we were able to derive asymptotics for this recursion, we could then figure out how well our α strategy does compared to the optimal and see if the α strategy is asymptotically competitive with optimal play. It would also interesting to explore for α strategies the idea of decreasing the threshold as we progress in the sequence. Intuitively, it makes sense the further along we get in a sequence to decrease α, as the probability of finding a bandit in the remaining ones that beats the bandit you're currently on decreases. We did not analyze these types of strategies as the analysis is complicated by the fact that α is dependent on where you are in the sequence. It would be interesting to see how these strategies perform compare to the strategies we explore in this paper.

Finally, we provide two very different families of strategies depending on how large K is relative to N. However, there is still a grey zone for $n^4 \leq K \leq n^6$ where we have no provably "good" strategies (although empirically, the strategy for large K works quite well for $K \geq N^4$). It would be interesting to explore this gap and where the theoretical threshold lies between switching from the family of strategies meant for K small relative to N to K large relative to N. Furthermore, are there any strategies that work equally well for both classes of K? We were unable to think of any during our exploration of this problem, but perhaps some family of strategies exist.

7 Acknowledgements

We would like to thank Professor Karger and Professor Madry for teaching this class and providing us with this great opportunity to conduct our own research in theoretical computer science. Further we'd like to thank Philip Sun for implementing an extremely optimized version of our 4 variable recursion for the optimal strategy in the case that $K = O(N)$ so that we would have an empirical benchmark to test any potential strategies against and also attempt to analyze heuristics and asymptotics of the optimal. We would also like to thank Patrick Insinger for running several of our intensive empirical computations, including computation of the optimal strategy recursion for large values of K, on his powerful desktop computer, without which we would have been unable to have such wealth of empirical data. We hope that our work has provided the initial foray into this variant of the multi-armed bandit problem and that future research can build off of our techniques and results.

References

[BMS09] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.

[Fin03] Finch. *Mathematical Constants.* 2003.

[KP14] Volodymyr Kuleshov and Doina Precup. Algorithms for multi-armed bandit problems. *arXiv preprint arXiv:1402.6028*, 2014.

[MT08] Aditya Mahajan and Demosthenis Teneketzis. Multi-armed bandit problems. In *Foundations and Applications of Sensor Management*, pages 121–151. Springer, 2008.

[OEI09] OEIS. Online encyclopedia of integer sequences, 2009.

[Sco15] Steven L. Scott. Multi-armed bandit experiments in the online service economy. *Applied Stochastic Models in Business and Industry*, 31:37–49, 2015. Special issue on actual impact and future perspectives on stochastic modelling in business and industry.