Bandit Algorithms

Divij Khaitan

December 8, 2024

1 Background: Reinforcement Learning

Reincforcement Learning comes from the theory of conditioning in psychology. Since it is one of the most popular theories of learning in humans, many have tried to replicate it to teaching machines to learn from data. Broadly speaking, reinforcement learning is process of placing an agent in an unknown environment, and having it learn to do certain tasks by having it interact with. To guide the learning process, Some key features of reinforcement learning include:

1. Learning by interaction: Humans often learn skills by watching others perform them and then trying constantly trying to replicate it until they get it right. Reinforcement learning uses this idea of an agent learning through interaction with it's immediate environment and being given a reward signal

2. No explicit supervision: While the above might sound like supervised learning, it has a subtle difference in that the agent is never told what the correct or optimal choice is. Much like when teaching a child how to ride a bicycle a parent will not specify which order specific muscles are relaxed and contracted. Similarly, the agent will have some reward signal attached to it's actions

3. Not Unsupervised: Reinforcement learning is also not completely unsupervised. The learner isn't explicitly trying to uncover some kind of substructure in a large dataset. Instead, it will be guided to and from certain sets of actions by the reward signal

4. Exploration-Explotation Dillema: While simply taking the most rewarding action at each step might seem like intuitive behaviour, it may not be globally optimal. Thus, an agent must learn to balance exploration of new actions with the exploitation of actions known to be rewarding

5. Limitations of Classical Methods: (Taken from [4]). Tic Tac Toe is a relatively simple problem that is 'solved' by methods like the minimax algorithm. However, even though the minimax solution never loses to a perfect player, it may not be optimal against an imperfect player. While

this could be remedied for simplistic situations, a game like chess would quickly go out of control.

The full reinforcement learning problem is fairly complex, and the multi-armed bandit problem is a common simplification that illustrates many of its ideas. The next section discusses the formal definition of the same.

2 The Bandit Problem

A bandit problem can be described using a tuple $(\mathcal{A}, R_i, \mathcal{B})$, where

- $\mathcal{A}$: The set of arms or actions, all the possible choices the agent can make in a particular turn. For the purposes of this paper, we will only be dealing with bandits that have a finite number of arms, so $|A| = k \in Z^+$
- R_i: The reward function is a map from $[T] -> \mathbb{R}$, and is defined for every arm
- $\mathcal{B}$: The discount factor. $\mathcal{B}_i$ is the effective value of the reward recieved at the i^{th} turn. This is primarily done to keep the expected reward from any action in an infinite horizon game finite.

The learner or agent is defined using a function $\pi(x|A_1, r_1, A_2, r_2, \ldots, A_{t-1}, r_{t-1})$, where the learner outputs a distribution over actions using the information contained by the game upto that point. This definition allows for a determinstic learner, by making the output distribution 0 for all the actions save for the action to be chosen. The game itself proceeds as follows

1. At time period t, the learner chooses action A_t in accordance with $\pi(x|A_1, r_1, A_2, r_2, \ldots, A_{t-1}, r_{t-1})$
2. The learner observes reward $R_{A_t}(t) = r_t$ 'chosen' by the environment and updates the $\pi \leftarrow \pi(x|A_1, r_1, A_2, r_2, \ldots, A_t, r_t)$

This repeats for a specified number of turns, finite or infinite. There are 2 distinct types of bandit problems we will encounter.

2.1 Stochastic Bandits

Stochastic Bandits are a collection of probability distributions $\nu = \{P_a | a \in \mathcal{A}\}$. In problems involving these bandits, the environment samples $r_(t) \sim P_{A_t}$. The reward distribution conditioned on $A_1, r_1, A_2, r_2, \ldots, A_{t-1}, r_{t-1}, A_t$ is exactly P_{A_t}, and learner cannot use any future information about their own actions or the rewards. The learner generally does not know the parameters of these distributions, and often not even the horizon of the game. These bandits can further be subdivided into two classes based on the information gained from playing each arm.

- Unstructured Bandits are problems where playing one arm gives you no information about the other arms. E.g. A two armed bandit with both arms having i.i.d. gaussian rewards.
- Structured Bandits are problems where playing an arm may give you some information about the other arm. E.g. The same two armed bandit as above with the additional stipulation that the means of both arms lie on the line $x + y = 2$

2.2 Adversarial Bandits

The Adversarial Bandit models abandons the assumption that the rewards from each arm are drawn from some distribution. The adversary chooses the rewards $x_t \in \mathbb{R}^{|\mathcal{A}|}$ for the entire duration of the game beforehand without the learner knowing.

A common relaxation in bandit problems, both stochastic and non-stochastic, is that the rewards lie in the interval [0, 1]. This simplifies the analysis significantly by allowing to use concentrations on subgaussian random variables and extends cleanly to complex cases.

2.3 Evaluation of Bandit Algorithms: Regret

The goal of this game is to maximise the total reward recieved. This isn't possible in the traditional sense for stochastic rewards, since an arm that is worse might produce a higher reward. It is better to define an objective function on each arm, and measure the distance of each decision an algorithm makes from the best arm. The most commonly used objective function is the mean of each arm. This is known as the regret, and is defined as

$$\text{Regret}_T = T \max_a \mu_a - \sum_{t=1}^{T} \mu(A_t)$$

Regret Decomposition Lemma:

$$\text{Regret}_T = \sum_{i=1}^{n} \Delta_i \mathbb{E}[n_i(T)]$$

3 ϵ greedy Algorithm

The epsilon greedy algorithm is the simplest algorithm for bandit problems. The pseudocode for the algorithm is as follows. The algorithm explores with probability ϵ, and exploits the best arm otherwise.

for $t = 0, 1 \ldots$ **do**
  Generate a random number $x \in [0, 1]$;
  if $i \leq \epsilon$ **then**
    Choose $A_t \sim \mathcal{U}([k])$;
  else
    Choose $A_t = \arg\max_a \hat{\mu_a}$;
  end
end

3.1 Analysis

This analysis was helped from [3]. For a bandit with N arms, at any iteration t the estimated mean for every arm a must meet the following bound (by hoeffding's inequality)

$$\mathbb{P}(|\hat{\mu_a} - \mu_a| \geq c) \leq 2\exp(-2c^2\frac{t\epsilon}{N})$$

Choosing $c = \sqrt{\frac{2N\log t}{t\epsilon}}$, we get

$$\mathbb{P}(|\hat{\mu_a} - \mu_a| \geq \sqrt{\frac{2N\log t}{t\epsilon}}) \leq \frac{2}{t^4}$$

We are largely concerned with bounding the regret of the rounds where the greedy arm is being exploited, since that having a bad arm would drive up regret. We can split the rounds into cases, one where the inequality holds for all the arms and the other where the inequality for at least one arm
In the first case, an arm being sampled implies that it has a higher estimated mean than the optimal arm. This gives us the inequality

$$\mu(a) + \sqrt{\frac{2N\log t}{t\epsilon}} \geq \hat{\mu}(a) \geq \hat{\mu}(a^*) \geq \mu(a^*) - \sqrt{\frac{2N\log t}{t\epsilon}}$$

$$\hat{\mu}(a^*) - \mu(a) \leq 2\sqrt{\frac{2N\log t}{t\epsilon}}$$

The expected regret in any single round can be expressed as

$$\epsilon(\text{Regret from a Random Arm}) + (1-\epsilon)(\text{Regret from largest sample mean})$$

$$\leq \epsilon + 2(1-\epsilon)\sqrt{\frac{2N\log t}{t\epsilon}}$$

This quantity can be minimised by roughly equating the two terms. This gives us an $\epsilon = (\frac{N\log t}{t})^{\frac{1}{3}}$ and a regret bound of

$$(\frac{N\log t}{t})^{\frac{1}{3}} + 2\sqrt{2}(1 - (\frac{N\log t}{t})^{\frac{1}{3}})(\frac{(N\log t)}{t})^{\frac{1}{3}}$$

This is an expression with $O((\frac{N\log t}{t})^{\frac{1}{3}})$. Thus, the total regret bound for the algorithm is $O(t^{\frac{2}{3}}(\frac{N\log t}{t})^{\frac{1}{3}})$

4 Upper Confidence Bound Algorithm

The idea behind the UCB algorithm is what is called the "OFU" principle - optimism in the face of uncertainty. The Upper Confidence Bound algorithm estimates the largest value that the mean of any particular arm could be with probability greater than some pre-defined Δ. The algorithm plays each arm once and subsequently plays the arm with the highest UCB in each round. Quantities attached to arms, called indexes are the bases for several bandit algorithms and these algorithms are called index algorithms.

for $t = 1 \dots$ **do**
  if $t \leq N$ **then**
    Choose $A_t = t$;
    Set $n_t = 1$;
    Set $\hat{\mu}_t = r_t$;
    Set $UCB_t = r_t + \sqrt{2\ln(t)}$;
  else
    Choose $A_t = \arg\max UCB_t$;
    Update $n_{A_t} += 1$;
    Update $\hat{\mu}_{A_t} = \hat{\mu}_{A_t} + \frac{r_t}{t}$;
    Update $UCB_{A_t} = \hat{\mu}_{A_t} + \sqrt{\frac{2\ln(t)}{n_{A_t}(t)}}$;
  end
end

4.1 Analysis

This analysis was taken from [2]. For any reward distribution in [0, 1], the UCB algorithm has regret which is upper bounded by

$$(8 \sum_{i:(\mu_{A_i} < \mu^*)} \frac{\log(n)}{\Delta_i}) + (1 + \frac{\pi^2}{3})(\sum_{j=1}^{k} \Delta_j)$$

Let $c_{t,s} = \sqrt{\frac{2\log(t)}{s}}$
We can bound the number of times $n_i(t)$ that arm i is played upto time t

$$n_i(T) = 1 + \sum_{t=N+1}^{T} I_{A_t = i}$$

Assuming the arm has been played l times so far, this quantity is never smaller than

$$\leq l + \sum_{t=N+1}^{T} \{A_t = i \ \wedge \ n_i(t) \geq l\}$$

$$\leq l + \sum_{t=N+1}^{T} \{(UCB^*(t) \leq UCB_{A_t}(t)) \wedge \ n_i(t) \geq l\}$$

$$\leq l + \sum_{t=N+1}^{T} \{(\hat{\mu}(a^*) + c_{(t-1),n_{a*}(t-1)} \leq \hat{\mu}(A_t) + c_{(t-1),n_{A_t}(t-1)}) \wedge \ n_i(t) \geq l\}$$

This has to be smaller than the number of times the smallest observed value of $UCB(a^*) \leq$ largest observed value of $UCB(A_t)$ upto a specified point s

$$\leq l + \sum_{t=N+1}^{T} \{((\min_{0 \leq s \leq t} \hat{\mu}_s(a^*) + c_{(t-1),s}) \leq (\max_{l \leq s_i \leq m} \hat{\mu}_{s_i}(A_t) + c_{(t-1),s_i}))\}$$

$$\leq l + \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{s_i=l}^{t-1} \{((\hat{\mu}_s(a^*) + c_{t,s}) \leq (\hat{\mu}_{s_i}(A_t) + c_{t,s_i}))\}$$

We can see that this is just the probability that a random arm with s_i plays has a higher UCB than an the optimal arm with s plays. This means that one of three conditions is true.

1. $\hat{\mu}_s(a^*) \leq \mu^* - c_{t,s}$
2. $\hat{\mu}_s(A_t) \geq \mu_i + c_{t,s}$
3. $\mu(a^*) \leq \mu(A_t) + 2c_{t,s_i}$

The probabilities of the first two situations can be bounded using the chernoff-hoeffding inequality for bounded random variables, as

$$\mathbb{P}(\hat{\mu}_s(a^*) \leq \mu(a^*) - \sqrt{\frac{2\log(t)}{s}}) \leq e^{-4ln(t)} = \frac{1}{t^4}$$

$$\mathbb{P}(\hat{\mu}_s(a^*) \leq \mu(a^*) - \sqrt{\frac{2\log(t)}{s}}) \leq e^{-4ln(t)} = \frac{1}{t^4}$$

The third situation cannot hold true for $l = \lceil \frac{8\log(t)}{\Delta_i^2} \rceil$. This can be shown as such

$$\mu(a^*) - \mu(A_t) - 2\sqrt{\frac{2\log(t)}{s_i}} \geq \mu(a^*) - \mu(A_t) - 2\sqrt{\frac{2\log(t)\Delta_i^2}{8\log(t)}}$$

$$\mu(a^*) - \mu(A_t) - 2\frac{\Delta_i}{2} = 0$$

This shows that $\mu(a^*) \geq \mu(A_t) + 2c_{t,s_i} \ \forall s_i \geq \lceil \frac{8\log(t)}{\Delta_i^2} \rceil$. Now, the number of plays can be bounded as

$$n_i(T) \leq \lceil \frac{8\log(t)}{\Delta_i^2} \rceil + \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{s_i = \frac{8\log(t)}{\Delta_i^2}}^{t-1} \frac{2}{t^{-4}}$$

The summand has no dependence on t for the two inner sums, so we can upper bound them by multiplying by t. This leaves us with a quadratic sum, which can be resolved by invoking the solution to the basel problem

$$\leq (\frac{8\log(t)}{\Delta_i^2} + 1) + 1 + \frac{\pi^2}{3}$$

$$\leq \frac{8\log(t)}{\Delta_i^2} + 1 + \frac{\pi^2}{3}$$

Using the regret decomposition lemma

$$\text{Regret}_T = \sum_{i=1}^{n} (\frac{8\log(t)}{\Delta_i^2} + 1 + \frac{\pi^2}{3})\Delta_i$$

$$\text{Regret}_T = \sum_{i:(\mu_{A_i} < \mu^*)} (\frac{8\log(t)}{\Delta_i}) + \sum_{i=1}^{n} (1 + \frac{\pi^2}{3})\Delta_i$$

5 Thompson Sampling

Thompson sampling is one of the oldest known algorithms for the bandit problem, first suggested in 1933 by William Thompson [5]. The algorithm models the reward distribution for each arm t as a conditional distribution $P_t(r|\theta_t)$. The nature of θ_t is dependent on the family of distributions that the arm comes from. For a gaussian bandit, θ_t may be the mean and variance of arm t and for a bernoulli bandit it might be the proportion of success. The main component of the thompson sampling is having a prior distribution associated with each arm, which is updated based on the samples seen from that arm. The idea is that exploration and exploitation are both balanced under this algorithm. Any tail event, such as an overestimate of a small arm or an underestimate of a large arm does not become very sub-optimal because the algorithm will draw more 'representative' sampled priors in the limit.

While experimentally known to be 'optimal', theoretical understanding of this algorithm was extremely limited. [1] was a seminal contribution on the subject that showed a logarithmic regret bound on the problem. For bernoulli bandits, TS is extremely straightforward. The priors are beta distributions initialised with $S, F = 0$. This is a natural choice of prior because the parameters can be used to track the number of successes and failures of the arm, and the concentration around the mean increases with increases to S or F. This can be seen in the density of the beta function, which is given by $\frac{\Gamma(S)}{\Gamma(S)\Gamma(F)} x^S (1-x)^F$. This has a mean of $\frac{S}{S+F}$. This can also conveniently be extended to the case of a bandit with rewards in the interval [0, 1] by adding an extra step to the algorithm - tossing a coin with probability r_t and updating the priors according to the outcome. This is helpful because for an arm with pdf f_i

$$P(r_t = 1) = \int_0^1 x f_i(x) dx = \mu_t$$

A way to interpret this is that this algorithm is estimating the mean of the distribution the unmodified algorithm would for a bernoulli bandit. This reduces the general case of rewards in [0, 1] down to solving the case for a bernoulli bandit.

S $= 0$, **F** $= 0$
for $t = 1 \ldots$ **do**
  Sample $\theta \sim \text{Beta}(S, F)$
  Choose $A_t = \arg\max_i \theta_i$
  Recieve reward r_t
  Update $S_{A_t} += Ir_t = 1$, $F_{A_t} += Ir_t = 0$
end

5.1 Analysis

The full proof is fairly involved, so a simplified version for the two arm case is given below. Note that the proof is only for a finite time horizon. Playing the optimal arm generates zero regret, and playing the second arm Δ regret. Call these arms (1) and (2). The number of plays of the (2) can be bounded as

$$E[n_2(T)] \leq L + \mathbb{E}[\sum_{j=j_0}^{T-1} Y_j]$$

Here, $L = \frac{24 \log(T)}{\Delta^2}$, j_0 is the number of plays of (1) until L plays of (2), Y_j is the number of times arm (2) is pulled in between the $j-1^{th}$ and j^{th} pulls of (1) To bound the Y_js, a random variable X is defined with parameters (j, s, y), which is like a geometric random variables counting the number of samples needed from Beta(s + 1, j - s + 1), before it exceeds some threshold y. It does not count the trial on which the experiment succeeds. This has expectation

$$\mathbb{E}[X(j, s, y)] = \frac{1}{1 - F^{\beta}_{s+1, j-s+1}} - 1 = \frac{1}{F^{bin}_{j+1, y}}(s)$$

The first equality follows from the fact that X is one less than a geometric random variable. The second is because $F^{\beta}_{a,b}(x) = 1 - F^{bin}_{a+b-1,x}(a-1)$. The Beta distribution has its CDF equal to the incomplete gamma function $I_x(a, b) = \frac{1}{\text{Beta}(a,b)} \int_0^x t^{a-1}(1-t)^{b-1} dt = 1 - F^{bin}_{a+b-1,x}(a-1)$.

This fact can be shown by the following. For an i.i.d. uniform sample $X_1, \ldots, x_{a+b-1}$ the ascending order statistics $X_{(a)} \sim \text{Beta}(a, b)$. This means that $F^{beta}_{a,b}(y)$ is the probability that $X_{(a)} \leq y$. Using the binomial distribution, $X_{(a)} \leq y$ iff at least a many samples are less than or equal to y. Since the X_is are uniform, $\mathbb{P}[X_i \leq y] = y$.

To show that the order statistics are beta distributed, the formula from [6], page 6 can be used.

$$f_{(k)}(x) = nf(x)^{n-1} C_{k-1} F^{k-1}(x)(1 - F(x))^{n-k}$$

for the cdf F and pdf f. For the uniform distribution, the cdf is the identity function while the pdf is the constant function 1. Using a sample of $a+b-1$, we get the a^{th} order statistic as

$$f_{(a)}(x) = (a+b-1)^{a+b-2} C_{a-1} x^{k-1} (1-x)^{n-k}$$

$$f_{(a)}(x) = \frac{(a+b-1)!}{(a-1)!(b-1)!} x^{k-1}(1-x)^{n-k}$$

The fraction at the start is exactly the normalisation constant $\frac{1}{\mathrm{B}(a,b)}$. This gives us

$$f_{(a)}(x) = F^{beta}_{a,b}(x)$$

The number of pulls after the j^{th} pull of (1) that the event $\{\theta_{(1)} \geq \mu_{(2)} + \frac{\Delta}{2}\}$ occurs is distributed as $X(j, s, \mu_{(2)} + \frac{\Delta}{2})$. Y_j can only be larger if at some point between these pulls $\theta_{(2)} \geq \mu_{(2)} + \frac{\Delta}{2}$, in which case $Y_j \leq T$. This gives the bound on the expectation

$$\mathbb{E}[Y_j] \leq \mathbb{E}[\min\{X(j, s(j), \mu_{(2)} + \frac{\Delta}{2}), T\}] + \mathbb{E}[\sum_{t=t_j+1}^{t_{j+1}-1} T\{\theta_{(2)} \geq \mu_{(2)} + \frac{\Delta}{2}\}]$$

This is splitting Y_j into two categories. The first is a 'good' case where $\theta_{(2)}$ is never larger than $\mu_{(2)} + \frac{\Delta}{2}$ and the second is the bad case where there is at least sample of θ for which $\theta_{(2)} \geq \mu_{(2)} + \frac{\Delta}{2}$. In this case, we bound the number of plays by T. This bound can be extended for any $j \geq j_0$

$$\mathbb{E}[Y_j\{j \geq j_0\}] \leq \mathbb{E}[\min\{X(j, s(j), \mu_{(2)} + \frac{\Delta}{2}), T\}] + \mathbb{E}[\sum_{t=t_j+1}^{t_{j+1}-1} T\{(\theta_{(2)} \geq \mu_{(2)} + \frac{\Delta}{2}) \wedge (j \geq j_0)\}]$$

Taking the sum over all possible j values

$$\sum_{j=j_0}^{T-1} \mathbb{E}[Y_j\{j \geq j_0\}] \leq \sum_{j=0}^{T-1} (\mathbb{E}[\min\{X(j, s(j), \mu_{(2)} + \frac{\Delta}{2}), T\}] + T\mathbb{E}[\sum_{t=t_j+1}^{t_{j+1}-1} \{(\theta_{(2)} \geq \mu_{(2)} + \frac{\Delta}{2}) \wedge (j \geq j_0)\}])$$

However, if $j \geq j_0$ then $n_{(2)}(t) \geq L$ which gives us

$$\leq \sum_{j=0}^{T-1} \mathbb{E}[\min\{X(j, s(j), \mu_{(2)} + \frac{\Delta}{2}), T\}] + T \sum_{t=1}^{T-1} \mathbb{P}((\theta_{(2)}(t) \geq \mu_2 + \frac{\Delta}{2}) \wedge (n_2(t) \geq L))$$

Let $M = (\theta_{(2)}(t) \geq \mu_{(2)} + \frac{\Delta}{2}) \wedge (n_2(t) \geq L)$. This can be seen as the tail event that despite many plays of (2), $\theta_{(2)}$ is much larger than $\mu_{(2)}$. We can bound this quantity.

Let $s_{(2)}(t)$ be the number of successes until the t^{th} step, $\hat{\mu}_(i) = \frac{s_{(i)}(t)}{n_{(i)}(t)}$ and define a random variable $A(t) = \{\hat{\mu}_{(2)} \leq \mu_{(2)} + \frac{\Delta}{4}\}$

$$\mathbb{P}(M) = \mathbb{P}(\theta_{(2)}(t) \geq \mu_2 + \frac{\Delta}{2} \wedge (n_2(t) \geq L))$$

$$\leq \mathbb{P}(\{A(t)=0\} \wedge (n_2(t) \geq L)) + \mathbb{P}(M \wedge \{A(t)=1\})$$

The first term is disjoint union of several binomial random variables with mean $\mu_{(2)}$. We can bound this using binomial chernoff bound

$$\leq \mathbb{P}(\{A(t)=0\} \wedge (n_2(t) \geq L)) = \sum_{l=L}^{T} \mathbb{P}(\{A(t)=0\} \wedge (n_2(t)=L))$$

$$\leq \sum_{l=L}^{T} \mathbb{P}(\{A(t)=0\})$$

$$\leq \sum_{l=L}^{T} \exp(\frac{-2l\Delta^2}{16}) \leq \frac{1}{T^2}$$

The second term will be bound in a similar way. We define $W_{l,z} \sim \text{Beta}(lz+1, l-lz+1)$. The number of plays of the suboptimal arm are distributed as $\text{Beta}(l\hat{\mu}_{(2)}+1, l-l\hat{\mu_{(2)}}+1)$.

$$\mathbb{P}(\{M\} \wedge \{A(t)=1\}) = \mathbb{P}(\{\theta_{(2)}(t) \geq \mu_{(2)} + \frac{\Delta}{2}\} \wedge (n_2(t)=l) \wedge \{A(t)=1\})$$

$$\leq \mathbb{P}(\{\theta_{(2)}(t) \geq \hat{\mu}_{(2)} + \frac{\Delta}{2} - \frac{\Delta}{4}\} \wedge (n_2(t)=l))$$

$$\leq \mathbb{P}(\{W(l, \hat{\mu}_{(2)}) \geq \hat{\mu}_{(2)} + \frac{\Delta}{4}\})$$

Rewriting using the relationship between the binomial and beta distributions gives us

$$= \sum_{l=L}^{T} \mathbb{E}[F^{bin}_{l+1,\hat{\mu}_{(2)}}(l\hat{\mu}_{(2)})]$$

Reducing the number of trials for a binomial random variable would increase the CDF at any single point.

$$\leq \sum_{l=L}^{T} \mathbb{E}[F^{bin}_{l,\hat{\mu}_{(2)}}(l\hat{\mu}_{(2)})]$$

We can now apply concentration bounds

$$\leq \sum_{l=L}^{T} \exp(-\frac{2\Delta^2 l}{16}) \leq T\exp(-\frac{2\Delta^2}{16}) \leq \frac{1}{T^2}$$

This gives us $\mathbb{P}(M) \leq \frac{2}{T^2}$. Plugging this back into the original equation gives

$$\sum_{j=j_0}^{T-1} \mathbb{E}[Y_j\{j \geq j_0\}] \leq \sum_{j=0}^{T-1} \mathbb{E}[\min\{X(j, s(j), \mu_{(2)} + \frac{\Delta}{2}), T\}] + T\sum_{t=1}^{T-1} \frac{2}{T^2}$$

$$\leq \sum_{j=0}^{T-1} \mathbb{E}[\min\{X(j, s(j), \mu_{(2)} + \frac{\Delta}{2}), T\}] + 2$$

To bound the first term, we can prove the following statement. For any $0 < y \leq \mu_{(1)}$, let $\Delta' = \mu_{(1)} - y$, $D = y\log(\frac{y}{\mu_{(1)}}) + (1-y)\log(\frac{(1-y)}{1-\mu_{(1)}})$ and $R = \frac{\mu_{(1)}(1-y)}{y(1-\mu_{(1)})}$

$$\mathbb{E}[\mathbb{E}[\min\{X(j, s(j), y), T\}|s(j)]] \leq \begin{cases} 1 + \frac{2}{1-y} + \frac{\mu_{(1)}}{\Delta'}\exp(-Dj) & \text{if } j \leq \frac{y}{D}\log(R), \\ 1 + (\frac{R^y}{1-y} + \frac{\mu_{(1)}}{\Delta'})\exp(-Dj) & \text{if } \frac{y}{D}\log(R) \leq j \leq \frac{4\log(T)}{\Delta'^2} \\ \frac{16}{T}, & \text{if } \frac{4\log(T)}{\Delta'^2} < j. \end{cases}$$

For the last case, concentration bounds can be invoked. For any $s \geq (y + \frac{\Delta'}{2})j$

$$F^{bin}_{j+1,y}(s) \geq F^{bin}_{j,y}(y + \frac{\Delta'}{2})j \geq 1 - \frac{\exp(2\Delta')}{\exp(\frac{j\Delta'^2}{2})} \geq 1 - \frac{\exp(2\Delta')}{T^2} \geq 1 - \frac{8}{T^2}$$

This shows that $\mathbb{E}[X(j+1, s, y)] \leq \frac{1}{1-\frac{8}{T^2}} - 1$. The probability that $s(j)$ takes a value smaller than $(y + \frac{\Delta'}{2})j$ is small, and the expectation term can be bounded by using the value T. The probability can be bounded by is

$$F^{bin}_{j+1,\mu_{(1)}}(yj + \frac{\Delta' j}{2}) \leq F^{bin}_{j+1,\mu_{(1)}}(\mu_{(1)}j + \frac{\Delta' j}{2}) \leq \exp(\frac{-j\Delta'}{2}) \leq \frac{1}{T^2} < \frac{8}{T^2}$$

The final expectation bound is

$$\mathbb{E}[\mathbb{E}[\min\{X(j, s(j), y), T\}|s(j)]] \leq (1 - \frac{8}{T^2})((1 - \frac{8}{T^2})^{-1} - 1) + \frac{8}{T^2}T \leq \frac{16}{T}$$

f^{bin} is the biomial pmf. For the first two cases, we use

$$\mathbb{E}[\mathbb{E}[\min\{X(j, s(j), y), T\}|s(j)]] = \mathbb{E}[\frac{1}{F^{bin}_{j+1,y}(s(j))}] = \sum_{s=0}^{j}(\frac{f^{bin}_{j,\mu_{(1)}}(s)}{F^{bin}_{j+1,y}(s)}) - 1$$

The median of a binomial distribution is its mean rounded up or down. This means for $s \geq \lceil y(j+1) \rceil$, $\sum_{s=\lceil y(j+1) \rceil}^{j}(\frac{f^{bin}_{j,\mu_{(1)}}(s)}{F^{bin}_{j+1,y}(s)}) \leq 2$.

For $s \leq \lfloor yj \rfloor$, the import facts are $F^{bin}_{j+1,y}(s) = (1-y)F^{bin}_{j,y}(s) + yF^{bin}_{j,y}(s-1) \geq (1-y)F^{bin}_{j,y}(s)$ and $F^{bin}_{j,y}(s) \geq f^{bin}_{j,y}(s)$

$$\sum_{s=0}^{\lfloor yj \rfloor}(\frac{f^{bin}_{j,\mu_{(1)}}(s)}{F^{bin}_{j+1,y}(s)}) \leq \sum_{s=0}^{\lfloor yj \rfloor}(\frac{f^{bin}_{j,\mu_{(1)}}(s)}{f^{bin}_{j+1,y}(s)})$$

$$= \sum_{s=0}^{\lfloor yj \rfloor} \frac{1}{1-y} \frac{\mu^s_{(1)}(1-\mu_{(1)})^{j-s}}{y^s(1-y)^{j-s}}$$

$$\sum_{s=0}^{\lfloor yj \rfloor} \frac{1}{1-y} R^s \frac{(1-\mu_{(1)})^j}{(1-y)^j}$$

$$= \frac{1}{1-y} \frac{R^{\lfloor yj \rfloor + 1}}{R-1} \frac{(1-\mu_{(1)})^j}{(1-y)^j}$$

Removing the -1 and expanding using the definition of R

$$\leq \frac{1}{1-y} \frac{R}{R-1} \frac{\mu_{(1)}^{yj}(1-\mu_{(1)})^{j-yj}}{y^{yj}(1-y)^{j-yj}}$$

$$= \frac{\mu_{(1)}}{\mu_{(1)}-y} \exp(-Dj) = \frac{\mu_{(1)}}{\Delta'} \exp(-Dj)$$

In cases where $\lfloor yj \rfloor < \lceil yj \rceil < \lceil y(j+1) \rceil$ both terms above miss the case of $y = \lceil yj \rceil$. However, here $\lceil yj \rceil \leq yj + y$

$$\frac{f^{bin}_{j,\mu_{(1)}}(s)}{F^{bin}_{j+1,y}(s)} \leq \frac{1}{(1-y)F^{bin}_{j,y}(s)} \leq \frac{2}{(1-y)}$$

An alternative bound for this case is

$$\frac{f^{bin}_{j,\mu_{(1)}}(s)}{F^{bin}_{j+1,y}(s)} \leq \frac{f^{bin}_{j,\mu_{(1)}}(s)}{(1-y)F^{bin}_{j,y}(s)} \leq \frac{f^{bin}_{j,\mu_{(1)}}(s)}{(1-y)f^{bin}_{j,y}(s)}$$

$$\leq \frac{R^s(1-\mu_{(1)})^j}{(1-y)(1-y)^j} \leq \frac{R^{yj+j}(1-\mu_{(1)})^j}{(1-y)(1-y)^j}$$

$$\leq \frac{R^y}{(1-y)} \exp(-Dj)$$

This completes the proof, since we have covered all 3 cases. We can finally bound the expected plays, using $y = \hat{\mu}_{(2)} + \frac{\Delta}{2}$ and $\Delta' = \frac{\Delta}{2}$ we get

$$n_2(T) \leq L + \sum_{j=0}^{T-1} \mathbb{E}[\mathbb{E}[\min\{X(j, s(j), \mu_{(2)} + \frac{\Delta}{2}), T\}|s(j)]] + 2$$

$\leq L + \frac{16\log(T)}{\Delta^2} + \frac{2y\log(R)}{D(1-y)} + \sum_{j=0}^{(16\log(T)/\Delta^2)-1}(\frac{2\mu_{(1)}}{\Delta}\exp(-Dj)) + \sum_{j=\frac{y}{D}\log(R)}^{(16\log(T)/\Delta^2)-1} \frac{(R^y\exp(-Dj))}{1-y} + \frac{16}{T}T + 2$

$\leq \frac{40\log(T)}{\Delta^2} + \frac{2y\log(R)}{D(1-y)} + \sum_{j=0}^{(16\log(T)/\Delta^2)-1}(\frac{2\mu_{(1)}}{\Delta}\exp(-Dj)) + \sum_{j=\frac{y}{D}\log(R)}^{(16\log(T)/\Delta^2)-1} \frac{(R^y\exp(-Dj))}{1-y} + 18$

We can reindex the second series and simplify it

$\leq \frac{40\log(T)}{\Delta^2} + 18 + \frac{2y\log(R)}{D(1-y)} + \sum_{j=0}^{(16\log(T)/\Delta^2)-1}(\frac{2\mu_{(1)}}{\Delta}\exp(-Dj)) + \sum_{j=0}^{(16\log(T)/\Delta^2)-1-\frac{y}{D}\log(R)} \frac{(\exp(-Dj))}{1-y}$

Combining the two series terms, we can upper bound them by extending the second series

$\leq \frac{40\log(T)}{\Delta^2} + 18 + \frac{4y\log(R)}{D\Delta} + \sum_{j=0}^{(16\log(T)/\Delta^2)-1}(\frac{2\mu_{(1)}+1}{\Delta}\exp(-Dj))$

Notice that $y\log(R) = y\log(\frac{\mu_{(1)}(1-y)}{(1-\mu_{(1)})y}) = y(\log(\frac{\mu_{(1)}}{y}) + \log(\frac{(1-y)}{(1-\mu_{(1)})}))$
$\leq \mu_{(1)} + \frac{y}{1-y}(D - y\log(\frac{y}{\mu_{(1)}})) \leq 1 + \frac{y}{1-y}(D + \mu_{(1)}) \leq \frac{2(D+1)}{\Delta}$

$\sum_0^k \exp(-Dj) \leq \frac{1}{1-e^{-D}} \leq \max\{\frac{2}{D}, \frac{e}{e-1}\} \leq \frac{2}{\min\{D,1\}}$

Plugging these bounds back in

$\leq \frac{40\log(T)}{\Delta^2} + 18 + \frac{8(D+1)}{D\Delta^2} + (\frac{4\mu_{(1)}+2}{\Delta\min\{D,1\}})$

Pinsker's inequality for bernoulli random variables is with proportions μ_1 and μ_2 is

$$|\mu_1 - \mu_2| \leq \sqrt{\frac{1}{2}D_{KL}(\mu_1, \mu_2)}$$

In the established notation, this becomes

$$2\Delta'^2 \leq D$$

$$\frac{\Delta^2}{2} \leq D$$

This gives us the bound
$\leq \frac{40\log(T)}{\Delta^2} + 18 + \frac{8}{\Delta^2} + (\frac{12}{\Delta^3}) + \frac{16}{\Delta^4}$
Since $\Delta \leq 1$, increasing the exponent in the denominator increases the sum
$\leq \frac{40\log(T)}{\Delta^2} + 18 + \frac{8}{\Delta^4} + (\frac{12}{\Delta^4}) + \frac{16}{\Delta^4}$
$\leq \frac{40\log(T)}{\Delta^2} + \frac{36}{\Delta^4} + 18$
$\in O(\log(T))$
This is asymptotically optimal for the stochastic bandit problem

References

[1] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem, 2012.

[2] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2–3):235–256, May 2002.

[3] Ioannis Panageas. Lecture notes: Gradient descent and beyond, n.d. Accessed: 2024-11-20.

[4] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018.

[5] William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[6] Duke University. Lecture 15: Order statistics, 2012. Accessed: 2024-12-08.