Stochastic Bandits

January 12, 2025

1 Concentration Inequalities

Theorem 1 (Chebyshev's Inequality)**.** *Let X be any random variable. Then for all $t > 0$,*

$$\Pr(|X - E[X]| \geq t) \leq \frac{Var(X)}{t^2}.$$

Theorem 2 (Markov inequality)**.** *Let X be any random variable that takes only non negative values. Then for any $c > 0$,*

$$\Pr(X \geq cE[X]) \leq 1/c.$$

Theorem 3 (Hoffding's inequality)**.** *Suppose $X_1, \ldots, X_n$ be independent bounded random variables such that for all $i \in [n]$, we have $a_i \leq X_i \leq b_i$ for all $i \in [n]$. Let $S_n = X_1 + \cdots + X_n$ and so $E[S_n] = \sum_i E[X_i]$.*

For any $t > 0$, we have

$$\Pr(|S_n - E[S_n]| \geq t) \leq 2e^{-\frac{2t^2}{\sum_i (b_i - a_i)^2}}$$

If $0 \leq X_i \leq 1$ for all $i \in [n]$ then we have,

$$\Pr(|S_n - E[S_n]| \geq t) \leq 2e^{-\frac{2t^2}{n}}$$

Let $\bar{X}_n = S_n/n = \frac{X_1 + \cdots + X_n}{n}$. So $E[\bar{X}_n] = \frac{\sum_i E[X_i]}{n}$ For any $t > 0$, we have

$$\Pr(|\bar{X}_n - E[\bar{X}_n]| \geq t) \leq 2e^{-\frac{2t^2 n^2}{\sum_i (b_i - a_i)^2}}$$

When for all i, $0 \leq X_i \leq 1$ then

$$\Pr(|\bar{X}_n - E[\bar{X}_n]| \geq t) \leq 2e^{-2t^2 n}$$

Remark. *Chebyshev's and Hoffding's inequalities works even for r.v. that can take negative values.*

2 Weak Law of Large Numbers

Let $X_1, \ldots, X_n$ be n iid (independent and identically distributed) random variables. Let $E[X] = \mu$ and $Var(X) = v$. Let $\bar{X}_n = \frac{\sum_i X_i}{n}$ be the empirical average. Note that $E[\bar{X}_n] = \mu$.

When $n \to \infty$ then the empirical average $\bar{X}_n$ converges to the true mean μ. We can see this from Chebyshev's inequality. As $X_1, \ldots, X_n$ are independent we have

$$Var(\bar{X}_n) = Var(\frac{\sum_i X_i}{n}) == \frac{nVar(X)}{n^2} = \frac{Var(X)}{n}.$$

So as $n \to \infty$, the variance of $\bar{X}_n$ tends to 0. So we have

$$\Pr(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{Var(X)}{n\epsilon^2} = \frac{v}{n\epsilon^2}$$

So for any fixed $\epsilon > 0$, for sufficiently large values of n (this value will depend on ϵ), the probability $\Pr(|\bar{X}_n - \mu| \geq \epsilon)$ will approach 0.

3 Stochastic Bandits

We start with the basic model of bandits, which is independent rewards. An algorithm has K possible arms to choose from, and there are T rounds, both K and T are known in advance (we will see later that one can relax the assumption that T should be known in advance). Each arm a is associated with a reward distribution D_a (pmf/pdf) which is unknown to the algorithm. The mean reward of the distribution D_a will be most relevant for us and is denoted by μ_a, i.e., $\mu_a = \underset{r \sim D_a}{E}[r]$ ($r \sim D_a$ denotes that r is sampled/drawn from the distribution D_a).

In each round $t \in T$:

1. algorithm picks an arm $a_t \in [K]$.
2. reward r_t is sampled from the distribution D_{a_t}.
3. algorithm receives the reward r_t.

The algorithm can be randomized, i.e., in any round it can fix a probability distribution over arms and can pick an arm from this distribution.

Again, it is important to note that $\mu_1, \ldots, \mu_K$ (and distributions $D_1, \ldots, D_K$) are unknown to the algorithm. The algorithm only sees the rewards (of the arms pulled by the algorithm).

We make the following assumptions:

1. rewards are bounded. For simplicty, we will assume rewards at any round will be in $[0, 1]$. So the means $\mu_1, \ldots, \mu_K$ all are in $[0, 1]$.
2. as we are in stochastic setting, we assume all drawn rewards are independent.

An important point to note that, regardless of the algorithm is deterministic or randomized, the arms pulled by the algorithm is a random variable. Because the algorithm's decision to pull any arm in some round will depend on past history of rewards observed by the algorithm. Since the received rewards are itself a random variable, the arms pulled are also random variable. This should become clear when we will see some algorithms.

Notations: Throughout the course, we will stick to the following notations. The set of arms will be denoted by $[K]$ and the total number of rounds will be denoted by T. The best arm is the arm a that has highest mean reward. We will use a^* for the best arm and μ^* for the mean reward of the best arm. That is $a^* = \arg\max_a \mu_a$ and $\mu^* = \mu_{a^*} = \max_a \mu_a$. For any suboptimal arm, we will use $\Delta_a := \mu^* - \mu_a$ for the gap of arm a. Finally, the arm pulled by the algorithm in tth round will be denoted by a_t.

Regret: How do we measure the performance of an algorithm? Recall that the goal of the algorithm is to maximize the sum of rewards received in all rounds.

One standard approach is to compare the algorithm performance with the best possible algorithm that knows the distributions D_a for all $a \in [K]$. Note that if means $\mu_1, \ldots, \mu_K$ are known then best strategy to maximize total expected rewards is to always pick the arm a^* to get total expected reward of $T\mu^*$. It makes sense to define the regret of an algorithm after T rounds as

$$R(T) = \mu^* T - \sum_{t=1}^{T} \mu_{a_t} = \sum_{t=1}^{T} (\mu^* - \mu_{a_t})$$

where a_t is the arm pulled by the algorithm in the t th round. Thus $R(T)$ is the regret incurred by the algorithm after T rounds of not knowing the means $\mu_1, \mu_2, \ldots, \mu_K$. Note that the regret $R(T)$ is a random variable. We are interested in the expected regret $E[R(T)]$ of the algorithm.

$$E[R(T)] = T\mu^* - E[\sum_{t=1}^{T} \mu_{a_t}]$$

The expectation in the above definition is taken over all the randomness, i.e., randomness over the draw of rewards from the distributions $D_1, \ldots, D_K$ and internal randomness of the algorithm (if the algorithm is randomized).

Remark. *1. By the definition only, $R(T)$ and $E[R(T)]$ of an algorithm depends on the problem-instance, i.e., means $\mu_1, \ldots, \mu_K$. Of course, we want an algorithm whose expected regret is small for all instances. In the previous line, 'for all problem instances' is important. For example, consider a stupid algorithm that always pulls arm 1. This algorithm will have regret 0 if the arm 1 has mean 1 and other arms have mean 0. But of course this algorithm will suffer for other instance such as if $\mu_1 = 0$ and $\mu_2 = 1$*

and will have regret of T. So we want an algorithm that performs well on all instances.

2. *If the regret is small then the algorithm's performance is close to best performance when the distributions are known. So we kind of learn the distributions.*

3. *If $r_1, \ldots, r_T$ are rewards received by the algorithm in rounds $1, 2, \ldots, T$ respectively then one can see that $E[\sum_{i=1}^{T} r_t] = E[\sum_{t=1}^{T} \mu_{a_t}]$. Hence the expected regret of an algorithm is the difference of expected total rewards of best strategy that knows $\mu_1, \ldots, \mu_K$ and the expected total rewards of the algorithm.*

Note: In some text books, $R(T)$ is called random-regret or realized regret and $E[R(T)]$ is called Regret. Its just a matter of convention, what names should we use. Often, from the context it will be clear whether we are talking about $R[T]$ or $E[R(T)]$.

Our goal is to design an algorithm whose expected regret $E[R(T)]$ is as small as possible for all problem-instances. Like a general discussion on algorithms, we will ignore constants and only consider Big-O dependence on T. Also, we assume $T \geq K$ as any reasonable algorithm will pull each arm atleast once. Note that since reward in any round is in $[0, 1]$, $R(T) \leq T$ always. We want an algorithms for which $R(T)$ grows sublinear in T, i.e., $\frac{R(T)}{T} \to 0$ as $T \to \infty$. In other words, average regret per round should be 0 when T is large (so we essentially we have learned the unknown distributions). Smaller the regret's dependence on T, faster the rate of convergence to 0 for the average regret per round.

3.1 Explore Then Commit Algorithm

This algorithm is very intuitive and perhaps we will first come up with this algorithm. If we know the means $\mu_1, \ldots, \mu_K$ then in each round the best algorithm will pick the arm a^* (recall notation, a^* has highest mean $\mu^* = \max_{a \in A}\{\mu_a\}$) to maximize total expected rewards. This algorithm first try each arm m times and finds the estimate $\hat{\mu}_a$ of the mean μ_a for every arm a. Thereafter, the algorithm will always pick an arm that maximizes the empirical mean $\hat{\mu}_a$ (we will call this arm a'). The value of m is set to $\frac{T^{2/3}(\log T)^{1/3}}{K^{2/3}}$ to optimze the expected regret.

We will now prove our first theorem of this course.

Theorem 4. *For any instance (i.e., for any values of unknowns $\mu_1, \ldots, \mu_K$), the expected regret of the algorithm ETC is $O(T^{2/3}(\log T)^{1/3}K^{1/3})$.*

Proof. Let $\epsilon = \sqrt{\frac{5 \ln T}{m}}$. For any arm a, let Bad_a be the event that $|\hat{\mu}_a - \mu_a| \geq \epsilon$. As $\hat{\mu}_a = \frac{\sum_{i=1}^{m} r_{ai}}{m}$ and all rewards are in $[0, 1]$, from Hoffding's inequality, we have

$$\Pr(Bad_a) = \Pr(|\hat{\mu}_a - \mu_a| \geq \epsilon) \leq \frac{2}{e^{2\epsilon^2 m}} = \frac{2}{T^{10}}.$$

```
Algorithm 1: Explore Then Commit
1. $m = \frac{T^{2/3}(\log T)^{1/3}}{K^{2/3}}$;
2. Try each arm $m$ times. Let for arm $a$, $r_{a1}, r_{a2}, \ldots, r_{am}$ be the
   rewards received.;
/* Line 2 is exploration/investment phase.  In this phase,
   the algorithm tries each arm so even the worst arms.  But
   we hope that we will be able to find near-optimal arm for
   remaining rounds.                                        */
3. For each arm $a$, set $\hat{\mu}_a$ as the average received reward for the arm $a$,
   i.e., we have $\hat{\mu}_a = \frac{\sum_{i \in [m]} r_{ia}}{m}$. Let $a'$ be the arm that maximizes $\hat{\mu}_a$,
   i.e., $a' = \arg\max_a \{\hat{\mu}_a\}$.;
/* We hope $a'$ is the near-optimal arm, i.e., $\mu^* - \mu_{a'}$ is very
   small.                                                   */
4. From the round $N \cdot K + 1$, always pick the arm $a'$.
```

Let $Bad = \cup_a Bad_a$ be the event that for some arm a, $|\hat{\mu}_a - \mu_a| \geq \epsilon$ holds. So $Good = Bad^c$ is the event that, for all arms a, we have $|\hat{\mu}_a - \mu_a| < \epsilon$.

By union bound, we have

$$\Pr(Bad) \leq \sum_a \Pr(Bad_a) \leq K \cdot \frac{2}{T^{10}} \leq \frac{2}{T^9}$$

(as we are assuming $T \geq K$ as any reasonable algorithm will try each arm at least once).

Also,

$$\Pr(Good) = 1 - \Pr(Bad) \geq 1 - \frac{2}{T^9}$$

As the $\Pr(Bad)$ is negligible (we have chosen $\epsilon = \sqrt{\frac{5 \ln T}{m}}$ to ensure that this happens), it will suffice for us to only bound the expeceted regret conditioned on event Good. Following calculation formally show this.

$$\begin{aligned} E[R(T)] &= \Pr(Bad)E[R(T)|Bad] + \Pr(Good)E[R(T)|Good] \\ &\leq \frac{2}{T^9} \cdot T + 1 \cdot E[R(T)|Good] \\ &\leq \frac{2}{T^8} + \cdot E[R(T)|Good] \end{aligned}$$

the second inequality comes from $E[R(T)|Bad] \leq T$ as all rewards are in $[0, 1]$.

Bounding $E[R(T)|Good]$**.** We will now assume event *Good*, i.e., for all arms a, we have $|\hat{\mu}_a - \mu_a| < \epsilon$. Recall that $R(T) = \sum_{t=1}^{T} \mu^* - \mu_{a_t}$.

The contribution to regret in the investment phase can be at most mK (assuming worst case contribution of 1 in each round). The contribution to the regret from $(mK+1)$st round till end is $(T-mK-1)(\mu^* - \mu_{a'})$ (recall that a' is the arm that maximizes $\hat{\mu}_a$ and is always chosen from $(mK+1)$st round) . We now claim that $(\mu^* - \mu_{a'}) < 2\epsilon$. For now let us assume this claim and bound the regret. We will prove the claim later.

$$\begin{aligned} R(T)|Good &\leq mK + (T-mK-1)2\epsilon \\ &\leq mK + 2T\epsilon \\ &= mK + 2T\sqrt{\frac{5\ln T}{m}} \end{aligned}$$

As m inreases mK increases and $2T\sqrt{\frac{5\ln T}{m}}$ decreases so the above quantity will be maximized when $mK = 2T\sqrt{\frac{5\ln T}{m}}$, i.e., when $m = (2\sqrt{5})^{2/3}\frac{T^{2/3}(\log T)^{1/3}}{K^{2/3}}$. Substituting this value of m we get $R(T)$ (conditioned on *Good*) equal to $O(T^{2/3}K^{1/3}(\log T)^{1/3})$. Obviously, we also have then
$E[R(T)|Good] = O(T^{2/3}K^{1/3}(\log T)^{1/3})$.

From our earlier calculation, $E[R(T)] = \frac{2}{T^8} + \cdot E[R(T)|Good] = O(T^{2/3}K^{1/3}(\log T)^{1/3})$. Now all it remains to prove the claim that $(\mu^* - \mu_{a'}) < 2\epsilon$. Note that $\hat{\mu}_{a'} \geq \hat{\mu}_{a^*} > \mu^* - \epsilon$ and also $\mu_{a'} > \hat{\mu}_{a'} - \epsilon$. Thus we have $\mu^* - \mu_{a'} < 2\epsilon$.

□