

# Stochastic Bandits

January 16, 2025

## 1 Concentration Inequalities

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

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

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

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

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

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

So as  $n \rightarrow \infty$ , 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  $\epsilon$ ), 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  $\mu_a$ , i.e.,  $\mu_a = E_{r \sim D_a}[r]$  ( $r \sim D_a$  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  $\mu_1, \dots, \mu_K$  (and distributions  $D_1, \dots, D_K$ ) 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 simplicity, we will assume rewards at any round will be in  $[0, 1]$ . So the means  $\mu_1, \dots, \mu_K$  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  $\mu^*$  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  $t$ th 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  $\mu_1, \dots, \mu_K$  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  $\mu_1, \mu_2, \dots, \mu_K$ . Note that the regret  $R(T)$  is a random variable. We are interested in the expected regret  $E[R(T)]$  of the algorithm.

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

The expectation in the above definition is taken over all the randomness, i.e., randomness over the draw of rewards from the distributions  $D_1, \dots, D_K$  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  $\mu_1, \dots, \mu_K$ . 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  $r_1, \dots, r_T$  are rewards received by the algorithm in rounds  $1, 2, \dots, T$  respectively then one can see that  $E[\sum_{i=1}^T r_i] = E[\sum_{i=1}^T \mu_{a_i}]$ . Hence the expected regret of an algorithm is the difference of expected total rewards of best strategy that knows  $\mu_1, \dots, \mu_K$  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.,  $\frac{R(T)}{T} \rightarrow 0$  as  $T \rightarrow \infty$ . 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  $\mu_1, \dots, \mu_K$  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  $\mu_a$  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 optimize 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  $\mu_1, \dots, \mu_K$ ), 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 Hoeffding'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  $a$ ,  $r_{a1}, r_{a2}, \dots, r_{am}$  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 expected regret conditioned on event Good. Following calculation formally show this.

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

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  $m$  increases  $mK$  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,  $E[R(T)] = \frac{2}{T^8} + E[R(T)|Good] = O(T^{2/3}K^{1/3}(\log T)^{1/3})$ . 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  $\mu^* - \mu_{a'} < 2\epsilon$ . □

## 4 Successive Elimination Algorithm

One drawback of ETC algorithm is that it will continue to explore an arm large number of times ( $m$  times) even if an arm's reward history might suggest to not pull this arm further. In Successive Elimination algorithm, we discontinue the arm forever once we have belief that the arm is not good. Below is the high level description of this algorithm.

---

**Algorithm 2:** Successive Elimination - High Level Description

---

- 1) Pull every arm once;
  - 2) If there is 'sufficient evidence' that some arm  $a$  is not a good arm then remove this arm;
- Repeat the above steps over the remaining arms
- 

Now to describe Successive Elimination fully, we just need to specify what 'sufficient evidence' is. For the same, we introduce some notations. For any arm  $a$  and round  $t$ , let  $n_a(t)$  be the number of times the arm  $a$  is pulled till the round  $t$ . Obviously, we have  $\sum_a n_a(t) = t$ . Further, let  $\hat{\mu}_a(t)$  be the empirical mean of received rewards from the arm  $a$  till round  $t$ . Formally, let  $r_{ai}$  be the reward received from the arm  $a$  on the  $i$ th pull. Then we have  $\hat{\mu}_a(t) = \frac{\sum_{i=1}^{n_a(t)} r_{ai}}{n_a(t)}$ . Let  $\epsilon_a(t) = \sqrt{\frac{5 \log T}{n_a(t)}}$ . Finally, let  $UCB_a(t) = \hat{\mu}_a(t) + \epsilon_a(t)$  and  $LCB_a(t) = \hat{\mu}_a(t) - \epsilon_a(t)$ .

Now we describe the 'sufficient evidence' which Successive Elimination employs. Recall the analysis of ETC algorithm. There we define an event *Good* (and show that it holds with high probability) and show that conditioned on *Good*,  $\hat{\mu}_a - \epsilon \leq \mu \leq \hat{\mu}_a + \epsilon$  where  $\epsilon = \sqrt{\frac{5 \log T}{m}}$ . Here also, we will define an event *Good* (and show it will hold with high probability) conditioned on which for all arm  $a$  and round  $t$ , we will have  $LCB_a(t) \leq \mu \leq UCB_a(t)$ . Now if at any time  $t$ , we have  $UCB_a(t) < LCB_{a'}(t)$  for some arms  $a$  and  $a'$  then we know that  $\mu_a < \mu_{a'}$ . Hence, it is not a good strategy to pull arm  $a$  in any subsequent rounds because we will be better off pulling arm  $a'$ . In other words, we can eliminate the arm  $a$  for future.

---

**Algorithm 3:** Successive Elimination

---

```

Activate all the arms;
while #-rounds <  $T$  do
    Pull each active arm once (and receive rewards);
    Deactivate all arms  $a$  such that there exists some another arm  $a'$ 
    with  $LCB_a < UCB_{a'}$ ;
end

```

---

**Theorem 5.** For all instances, i.e., for all values of  $\mu_1, \dots, \mu_K$ ,

$$E[R(T)] = O(\sqrt{KT \log T})$$

*Proof.* Let *Good* be the following event: for all arms  $a$  and for all rounds  $t$ , we have  $|\hat{\mu}_a(t) - \mu_a| < \epsilon_a(t)$  (that is,  $LCB_a(t) \leq \mu_a \leq UCB_a(t)$ ). We will prove that

$$\Pr(\text{Good}) \geq 1 - O\left(\frac{1}{T^8}\right)$$

For now, let us assume the above. Like ETC, it will suffice to bound  $E[R(T)|\text{Good}]$ .

$$\begin{aligned} E[R(T)] &= \Pr(\text{Bad})E[R(T)|\text{Bad}] + \Pr(\text{Good})E[R(T)|\text{Good}] \\ &\leq O\left(\frac{1}{T^8}\right) \cdot T + 1 \cdot E[R(T)|\text{Good}] \\ &\leq O\left(\frac{1}{T^7}\right) + E[R(T)|\text{Good}] \end{aligned}$$

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

**Bounding  $E[R(T)|\text{Good}]$**  Each calculation in this paragraph is conditioned on *Good*. Note that  $R(T) = \sum_t (\mu^* - \mu_{a_t}) = \sum_a n_a(T)(\mu^* - \mu_a)$ . If we show for each arm  $a$ ,  $(\mu^* - \mu_a) \leq (8\sqrt{\frac{5 \log T}{n_a(T)}})$  then we are done. This is because  $R(T) = \sum_a n_a(T)(\mu^* - \mu_a) \leq \sum_a n_a(T)(8\sqrt{\frac{5 \log T}{n_a(T)}}) \leq 8\sqrt{5n_a(T) \log T}$ . As  $\sqrt{x}$  is a concave function so we have  $\frac{\sum_a \sqrt{n_a(T)}}{K} \leq \sqrt{\frac{\sum_a n_a(T)}{K}} = \sqrt{T/K}$ . Thus  $R(T) \leq 8\sqrt{5KT \log T}$ . So it remains to show that for any arm  $a$ , we have  $(\mu^* - \mu_a) \leq 8\sqrt{\frac{5 \log T}{n_a(T)}}$ . For the sake of analysis, we will refer to the  $i$ th iteration of while loop as phase  $i$  (for any  $i$ ). Our first easy observation is that the arm  $a^*$  with will never be deactivated. Let  $t$  be the last round (corresponding to the end of some phase) where the arm  $a$  remained active (in other words, arm  $a$  was played exactly once after  $t$ ). Note that  $n_a(t) = n_{a^*}(t)$  (because both arms  $a$  and  $a^*$  are active till  $t$  so both of them are played equal number of times, which is the number of phases completed till  $t$ ). As the arm  $a$  is not deactivated at  $t$ , we must have  $UCB_a(t) \geq LCB_{a^*}(t)$ . Further, we have  $\epsilon_a(t) = \epsilon_{a^*}(t) = \sqrt{\frac{5 \log T}{n_a(t)}} = \sqrt{\frac{5 \log T}{n_a(T)-1}}$ . It is easy to now see that we have  $\mu^* - \mu_a \leq 4\epsilon_a(t) = 4\sqrt{\frac{5 \log T}{n_a(T)-1}} \leq 8\sqrt{\frac{5 \log T}{n_a(T)}}$ .

**Bounding  $\Pr(\text{Good})$**  It remains to show :

$$\Pr(\text{Good}) \geq 1 - O\left(\frac{1}{T^8}\right)$$

Let  $r_{a1}, r_{a2}, \dots, r_{aT}$  be  $T$  samples from the distribution  $D_a$ . Let  $\text{Bad}_a(t)$  be the event that  $|\frac{r_{a1} + \dots + r_{at}}{t} - \mu_a| \geq \sqrt{\frac{5 \log T}{t}}$ . By Hoeffding's inequality, we have



$\Pr(Bad_a(t)) \leq O(1/e^{10 \log T}) = O(\frac{1}{T^{10}})$ . Let  $Bad_a$  be the event that for some  $1 \leq t \leq T$ , we have  $|\frac{r_{a1} + \dots + r_{at}}{t} - \mu_a| \geq \sqrt{\frac{5 \log T}{t}}$ . By union bound,  $\Pr(Bad_a) \leq T \cdot O(\frac{1}{T^{10}}) = O(\frac{1}{T^9})$ . Let  $Bad = \cup_a Bad_a$ . Again by union bound,  $\Pr(Bad) \leq K \cdot O(\frac{1}{T^9}) = O(\frac{1}{T^8})$ .  $\square$

The regret in the above theorem is worst-case regret, i.e, for any problem-instance (that is, for any values of  $K, T$  and  $\mu_1, \dots, \mu_K$ , the expected regret  $E[R(T)] \leq O(\sqrt{KT \log T})$ . Now we will show another type of bounds on expected regret, which will be instance -dependent bound.

**Theorem 6.** *The expected regret of Successive Elimination satisfies*

$$E[R(T)] \leq O(\log T) \sum_{a: \Delta_a > 0} \frac{1}{\Delta_a}$$

where  $\Delta_a = \mu^* - \mu_a$  is the gap of arm  $a$ .

*Proof.* We define the events  $Bad$  and  $Good$  as in the above theorem. Again it will suffice to bound  $E[R(T)|Good]$ . All calculations now are conditioned on  $Good$ . We claim that for any suboptimal arm  $a$ , we have  $n_a(T) \leq 50000 \frac{\log T}{\Delta_a^2}$ . This implies that  $R(T) = \sum_a n_a(T) \Delta_a \leq O(\log T) \sum_{a: \Delta_a > 0} \frac{1}{\Delta_a}$ . Suppose  $n_a(T) > 50000 \frac{\log T}{\Delta_a^2}$ . Consider the time  $t$  when  $n_a(t) = 50000 \frac{\log T}{\Delta_a^2}$ . As the arm  $a^*$  is always active we also have  $n_{a^*}(t) = 50000 \frac{\log T}{\Delta_a^2}$ . So now we have  $\epsilon_a(t) = \epsilon_{a^*}(t) = \sqrt{\frac{5 \log T}{n_a(t)}} = \Delta_a/100$ . As we have assumed  $Good$ , we have  $LCB_{a^*}(t) \geq \mu^* - \Delta_a/100$  and  $UCB_a(t) \leq \mu_a + \Delta_a/100$ . So we have  $UCB_a(t) < LCB_{a^*}(t)$  which implies that arm  $a$  will be eliminated in round  $t$  which contradicts that  $n_a(T) > 50000 \frac{\log T}{\Delta_a^2}$ .  $\square$