# Unimodal Stochastic Multi-Armed Bandit Problems

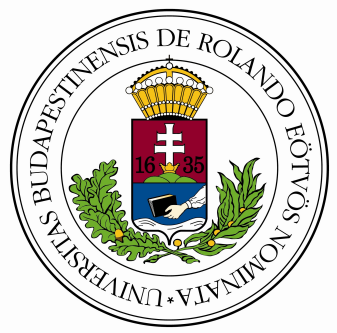
**Roland Szögi**

MSc in Applied Mathematics

Supervisor:

**Balázs Csanád Csáji, PhD**

Department of Probability Theory and Statistics, Eötvös Loránd University; Institute for Computer Science and Control, Hungarian Research Network (HUN-REN SZTAKI)




Eötvös Loránd University
Institute of Mathematics
Faculty of Science
Budapest, 2025

# Contents

# Acknowledgement

I would like to thank Balázs Csáji for his help and guidance.

# 1 Introduction

## 1.1 Overview of Known Results

The multi-armed bandit problem is one of the most studied problems in control and decision theory, which encompasses the fundamental dilemma between exploration and exploitation [1]. The term multi-armed bandit originates from the name of the slot machine called one-armed bandit, because the studied problem is similar to the situation a gambler might face in a casino when he wants to decide which machines to play from a row of slot machines. Each of these machines is associated with a probability distribution and pulling the arm of a machine yields a reward from its distribution. Initially the distributions are unknown, but as the gambler plays the machines, he gathers information which can guide future decisions. The gambler may have different objectives, such as maximizing the total reward over a fixed number of trials or identifying a near-optimal machine.

There are well-known algorithms to achieve these goals. The Upper Confidence Bound (UCB) algorithm can be used if the goal is to maximize the total reward. In this case we use the regret to measure how good an algorithm is. The regret is the difference between the maximal reward, that can be achieve over $n$ rounds by playing the optimal arm all the time and the expectation of the sum of the rewards collected by the algorithm. The smaller the regret, the better the algorithm. It can be shown, that the Upper Confidence Bound algorithm is asymptotically optimal, meaning that as the number of rounds tend to infinity, the regret becomes a constant multiple of the regret of the best possible algorithm.

If the goal is to identify a close-to-optimal arm, we can use the Median Elimination algorithm. In each round of this algorithm, we sample each remaining arm a specified number of times, calculate the sample mean for each arm, and eliminate the half of the arms with the lowest sample means.

Of course, for these algorithms to work, we must make some assumptions about the arms. We assume that they belong to a defined family of probability distributions. One of the most common assumptions is that the arms are 1-subgaussian, meaning that the tails of the distributions decay at least as quickly

as those of the standard normal distribution.

These algorithms can be used not only in casinos, the multi-armed bandit problem has many applications including A/B testing, advert placement and recommendation services [2]. In online advertising, multi-armed bandit algorithms can dynamically adjust ads to maximize click through rates by balancing exploration and exploitation [3, 4]. Multi-armed bandits power recommendation systems too, such as those used by streaming services or e-commerce platforms, to dynamically recommend products or content while learning user preferences over time [5, 6, 7]. They can also be used in adaptive clinical trials to dynamically allocate patients to treatments, aiming to find the most effective treatments while minimizing exposure to less effective ones [8, 9]. Bandit algorithms are applied in financial portfolio management to dynamically allocate assets in a way that maximizes returns while managing risk through diversification and adaptive learning of asset performance [10, 11]. In e-commerce and retail, these algorithms help in setting optimal prices for products by balancing the exploration of different pricing strategies and the exploitation of known profitable pricing [12].

In my thesis I present some of the principal results related to the stochastic multi-armed bandit problems, focusing on the Upper Confidence Bound and Median Elimination algorithms.

## 1.2 Contributions of the Thesis

In many applications, we can use the Median Elimination Algorithm to find a close-to-optimal arm, since this algorithm doesn't assume anything about the structure of the arms. However, in some applications we can assume that an unknown concave function describes the expectations of the arms and in this case we might use this information to find a close-to-optimal arm much faster.

For example, consider a company aiming to determine the optimal price to charge for a product in order to maximize its profit. In this scenario, each possible price point can be viewed as an arm in the multi-armed bandit problem. Initially, reducing the price from a higher value often leads to a substantial increase in the number of units sold, as more customers are attracted by the lower cost. However, after reaching

a certain threshold, further price reductions yield diminishing returns in terms of sales. This happens because customers become more price-sensitive, and after a point, additional reductions in price no longer correspond to proportional increases in demand. Instead, they might simply reduce the perceived value of the product. This way we can assume that the company's profit is a concave function of the price. In this case, the number of arms is infinite because the price is a continuous variable. However, if the company can only choose from a limited number of pricing strategies or variations (e.g., different levels of product quality), the problem becomes finite, as there is a fixed number of possible arms to test.

In my thesis, I examine some special cases of the multi-armed bandit problem and present my own algorithms which leverage the extra information to identify close-to-optimal arms more efficiently than Median Elimination. In all of these cases I assumed that the arms are 1-subgaussian. I have developed $(\varepsilon, \delta) - PAC$ algorithms for the following cases:

- There are finitely many arms, and we know that the expectations of the arms increase up to a certain arm, after which the expectation decreases, and the difference between the expectations of neighboring arms is at least $\Delta$, where $\Delta$ is known in advance. In this case the provided new algorithm finds the optimal arm much faster than the Median Elimination algorithm (by choosing $\varepsilon < \Delta$) with a sample complexity of

$$\mathcal{O}\left(\log n + \frac{1}{\Delta^2} \log \frac{n}{\delta}\right)$$

  instead of

$$\mathcal{O}\left(\frac{n}{\Delta^2} \log \frac{1}{\delta}\right).$$

- The arms are the points of the $[0, 1]$ interval and an unknown concave function describes the expectations of the arms.

- The arms are the points of the $[0, 1]$ interval and an unknown concave function describes the expectations of the arms, which is Lipschitz continuous with a known Lipschitz constant $L$. I have also calculated the sample complexity of the algorithm in this case.

- There are finitely many arms and an unknown concave function describes the expectations of the arms. The algorithm I developed finds an $\varepsilon$-optimal arm with probability at least $1 - \delta$ more efficiently than Median Elimination, with a sample complexity of

  $$\mathcal{O}\left(\frac{1}{\varepsilon^2}\left((\log n)^2 + \log n \cdot \log\frac{1}{\delta}\right)\right)$$

  instead of

  $$\mathcal{O}\left(\frac{n}{\varepsilon^2}\log\frac{1}{\delta}\right).$$

I have also implemented the algorithms and conducted experiments to assess their performance by plotting the relationship between the algorithm inputs and the sample complexity.

# 2 Definitions

We start this section by giving a more formal definition of the bandit problem. The definitions and results in this and the following two sections can be found in [2].

## 2.1 Stochastic Bandits

The bandit problem is a sequential game between a learner and an environment. The stochastic bandit model consists of a set of distributions $\nu = (\mathbb{P}_a : a \in \mathcal{A})$, where $\mathcal{A}$ is the set of available actions, which are often called arms in the literature. In each round $t$ the learner chooses an action $A_t \in \mathcal{A}$ and a reward $X_t$ is sampled from distribution $\mathbb{P}_{A_t}$ by the environment. Both the learner and the environment may use randomization.

The expectation of arm $a$ will be denoted by $\mu_a(\nu)$ and the largest expectation will be denoted by $\mu^*(\nu) = \max_{a \in \mathcal{A}} \mu_a(\nu)$. We assume throughout that $\mu_a(\nu)$ exists and that it is finite for all $a \in \mathcal{A}$. We also assume that $\mu^*(\nu)$ exists. If the context is clear, we will drop the dependence on $\nu$ and write $\mu_a$ instead of $\mu_a(\nu)$ and $\mu^*$ instead of $\mu^*(\nu)$.

In each round, the learner can only use the past observations to make a decision, so $A_t$ only depends on the history $H_{t-1} = (A_1, X_1, ..., A_{t-1}, X_{t-1})$. A policy is a mapping from histories to actions, which defines the decisions made by the learner based on the previous observations. Policies can use randomization when choosing the next action to be sampled and when making the final selection. An environment is a mapping, which maps histories ending in actions to the reward received by the learner. The environment is unknown to the learner, the learner only knows, that the true environment is in a given set of environments, called the environment set. The term bandit is often used instead of environment. For example an environment is called a stochastic Bernoulli bandit, if the rewards are binary and the distributions are Bernoulli distributions. The class of stochastic Bernoulli bandits is the set including all stochastic Bernoulli bandits characterized by their mean vectors.

We distinguish between structured and unstructured bandits. An environment $\mathcal{E}$ is called unstructured, if $\mathcal{A}$ is finite and there exist a set of distribution $\mathcal{M}_a$ for all

$a \in \mathcal{A}$ such that $\mathcal{E}$ is the Cartesian product of these sets

$$\mathcal{E} = \times_{a \in \mathcal{A}} \mathcal{M}_a.$$

In this case by playing action $a$ the leaner cannot learn anything about the distribution of arm $b \neq a$. One example of unstructured bandits is the Bernoulli bandit defined above.

Environment classes that are not unstructured are called structured. One simple example is the stochastic linear bandit, where $\mathcal{A} \subset \mathbb{R}^d, \theta \in \mathbb{R}^d$ and $\nu_\theta = (\mathcal{N}(\langle a, \theta \rangle, 1) : a \in \mathcal{A})$, $\mathcal{E} = \left\{\nu_\theta : \theta \in \mathbb{R}^d\right\}$. Here the learner can deduce the true environment by playing just $d$ arms spanning $\mathbb{R}^d$.

## 2.2 Subgaussian Random Variables

One of the most important and most often used model for unstructured bandit problems is the subgaussian bandit, where the distributions are subgaussian. In this section, we give the definition of subgaussian random variables and prove theorems about them, which will be used in later chapters.

**Definition 1.** *A random variable $X$ is $\sigma$-subgaussian if for all $\lambda \in \mathbb{R}$ :*

$$\mathbb{E}\left[\exp(\lambda X)\right] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2}\right).$$

A few examples of subgaussian random variables:

- If $X$ is Gaussian with mean zero and variance $\sigma^2$, then $X$ is $\sigma$-subgaussian.
- If $X$ has mean zero and $|X| \leq B$, then $X$ is $B$-subgaussian.
- If $X$ has mean zero and $X \in [a, b]$, then $X$ is $(b-a)/2$-subgaussian.

**Remark 1.** *For random variables that are not centered ($\mathbb{E}[X] \neq 0$), the notation is abused by saying that $X$ is $\sigma$-subgaussian if the noise $X - \mathbb{E}[X]$ is $\sigma$-subgaussian. A distribution is called $\sigma$-subgaussian if a random variable drawn from that distribution is $\sigma$-subgaussian.*

The following theorem states that the tails of a $\sigma$-subgaussian distribution decay at least as fast as the tails of a Gaussian distribution with zero mean and $\sigma$ standard deviation. This property gives subgaussian distributions their name.

**Theorem 1.** *If $X$ is $\sigma$-subgaussian, then for any $\varepsilon \geq 0$*

$$\mathbb{P}(X \geq \varepsilon) \leq \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right).$$

*Proof.* Based on Markov's inequality and the definition of subgaussianity, for all $\lambda > 0$:

$$\begin{aligned}\mathbb{P}(X \geq \varepsilon) &= \mathbb{P}\left(\exp(\lambda X) \geq \exp(\lambda\varepsilon)\right) \\ &\leq \mathbb{E}[\exp(\lambda X)]\exp(-\lambda\varepsilon) \\ &\leq \exp\left(\frac{\lambda^2\sigma^2}{2} - \lambda\varepsilon\right).\end{aligned}$$

By choosing $\lambda = \varepsilon/\sigma^2$ we get that

$$\mathbb{P}(X \geq \varepsilon) \leq \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right).$$

□

To study the tail behaviour of $\hat{\mu} - \mu$ we will use the following statements, which easily follow from the definition of subgaussian random variables.

**Statement 1.** *Assume that $X$ is $\sigma$-subgaussian and $X_1$ and $X_2$ are independent $\sigma_1$ and $\sigma_2$-subgaussian random variables. In this case:*

- *$cX$ is $|c|\sigma$-subgaussian for all $c \in \mathbb{R}$.*
- *$X_1 + X_2$ is $\sqrt{\sigma_1^2 + \sigma_2^2}$-subgaussian.*

The following theorem provides concentration bounds for the sample mean $\hat{\mu}$ of independent, $\sigma$-subgaussian random variables. It characterizes the probability that the sample mean deviates from the true mean $\mu$ by more than a specified amount $\varepsilon$.

**Theorem 2.** *Assume that $X_i - \mu$ are independent, $\sigma$-subgaussian random variables. Then for any $\varepsilon > 0$,*

$$\mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right),$$
$$\mathbb{P}(\hat{\mu} \leq \mu - \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right)$$

*where $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$.*

*Proof.* It follows from Statement 1, that $\hat{\mu} - \mu = \sum_{i=1}^n (X_i - \mu)/n$ is $\sigma/\sqrt{n}$-subgaussian. Using Theorem 1 for $\hat{\mu} - \mu$, we get the statement. □

The above result says that for any $\delta \in (0, 1)$ with probability at least $1 - \delta$,

$$\mu \leq \hat{\mu} + \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}.$$

# 3 Upper Confidence Bound Algorithm

In this section we will examine the case, when the goal is to maximize the expected cumulative reward, $\mathbb{E}[S_n] = \mathbb{E}\left[\sum_{t=1}^{n} X_t\right]$, over $n$ rounds.

## 3.1 Regret

The regret is an important measure of performance when the goal is to maximize the expected cumulative reward. The regret of policy $\pi$ on bandit $\nu$ is defined by

$$R_n(\pi, \nu) = n\mu^*(\nu) - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right],$$

where the expectation is taken with respect to the probability measure on outcomes induced by the interaction of $\pi$ and $\nu$. Minimizing the regret is equivalent to maximizing the expectation of the sum of rewards. If the context is clear, we will use $R_n$ instead of $R_n(\pi, \nu)$.

Based on the following statement, the regret is always non-negative, and for every bandit $\nu$, there exists a policy $\pi$ for which the regret is 0. Achieving zero regret is possible if and only if the learner chooses the optimal action in each round.

**Statement 2.** *Let $\nu$ be a stochastic bandit. Then,*

- *$R_n(\pi, \nu) \geq 0$ for all policies $\pi$,*
- *the policy $\pi$ choosing $A_t \in \arg\max \nu_a$ for all $t$ satisfies $R_n(\pi, \nu)$,*
- *if $R_n(\pi, \nu) = 0$ for policy $\pi$, then $\mathbb{P}(\mu_{A_t} = \mu^*) = 1$ for all $t$.*

The quantity $\Delta_a(\nu) = \mu^*(\nu) - \mu_a(\nu)$ is called the suboptimality gap, action gap or immediate regret of arm $a$. Let $T_a(t) = \sum_{s=1}^{t} \mathbb{I}\{A_s = a\}$ be the number of times action $a$ was sampled by the end of round $t$. The following lemma decomposes the regret in terms of the loss due to using each of the arms.

**Lemma 1.** *(Regret decomposition lemma)*
*Suppose that the set of actions $\mathcal{A}$ is finite or countable. Then for any policy $\pi$, bandit $\nu$ and horizon $n \in \mathbb{N}$, the regret of policy $\pi$ satisfies*

$$R_n = \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}\left[T_a(n)\right].$$

*Proof.* Since $\sum_{a\in\mathcal{A}} \mathbb{I}\{A_t = a\} = 1$ for any fixed $t$,

$$S_n = \sum_t X_t = \sum_t \sum_a X_t \mathbb{I}\{A_t = a\},$$

$$R_n = n\mu^* - \mathbb{E}[S_n] = \sum_{a\in\mathcal{A}} \sum_{t=1}^{n} \mathbb{E}\left[(\mu^* - X_t)\mathbb{I}\{A_t = a\}\right].$$

The expected reward in round $t$ conditioned on $A_t$ is $\mu_{A_t}$, this way

$$\begin{aligned}
\mathbb{E}\left[(\mu^* - X_t)\mathbb{I}\{A_t = a\}|A_t\right] &= \mathbb{I}\{A_t = a\}\mathbb{E}[\mu^* - X_t|A_t] \\
&= \mathbb{I}\{A_t = a\}(\mu^* - \mu_{A_t}) \\
&= \mathbb{I}\{A_t = a\}(\mu^* - \mu_a) \\
&= \mathbb{I}\{A_t = a\}\Delta a.
\end{aligned}$$

By substituting this into the equation of $R_n$ and using the definition of $T_a(n)$, we get the result. □

## 3.2 Algorithm

The Upper Confidence Bound (UCB) algorithm follows the optimism in the face of uncertainty principle, which means that based on the observed data each arm is assigned a value called upper confidence bound, which overestimates the unknown mean of the arm with high probability. An arm is played only if its upper confidence bound is larger than the upper confidence bound of the optimal arm. By playing a suboptimal arm, the upper confidence bound will fall below the upper confidence bound of the optimal arm. In the algorithm we assume that the arms are 1-subgaussian and that the arms are numbered from 1 to $k$.

If $(X_t)_{t=1}^n$ is a sequence of independent 1-subgaussian random variables with mean $\mu$ and sample mean $\hat{\mu} = \frac{1}{n}\sum_{t=1}^n X_t$, by Theorem 2, for all $\delta \in (0,1)$

$$\mathbb{P}\left(\mu \geq \hat{\mu} + \sqrt{\frac{2\log(1/\delta)}{n}}\right) \leq \delta.$$

We will use $T_i(t-1)$ to denote the number of times arm $i$ was sampled by round $t$ and $\hat{\mu}_i(t-1)$ will denote the sample mean of the obtained rewards. The upper confidence bound in this case is defined the following way

$$UCB_i(t-1,\delta) = \begin{cases} \infty, & \text{if } T_i(t-1) = 0 \\ \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}}, & \text{otherwise} \end{cases}$$

Here $\delta$ is called the confidence level and it quantifies the degree of certainty.

The intuition behind the algorithm is that the algorithm should explore promising arms ($\hat{\mu}_i(t-1)$ is large) and arms which are not well explored ($T_i(t-1)$ is small). Suppose that at the start of round $t$ the first arm has been sampled much more frequently than the rest and because of it we expect that $\hat{\mu}_1(t-1) \approx \mu_1$. In this case the learner can be certain that arm $i$ is worse than arm 1 if the upper confidence bound is less than or equal to the mean of arm 1:

$$\hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}} \leq \mu_1 \approx \hat{\mu}_1(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_1(t-1)}}.$$

By choosing the arm with the largest upper confidence bound, an arm is chosen only if its true mean could be larger than the mean of the arms that have been played often.

## 3.3 Bounding the Regret of UCB

We will provide two bounds on the regret of the UCB algorithm.

**Theorem 3.** *In case of the $k$-armed 1-subgaussian bandit problem, for any horizon $n$, if $\delta = 1/n^2$, then for the regret of the UCB algorithm*

$$R_n \leq 3\sum_{i=1}^{k} \Delta_i + \sum_{i:\Delta_i > 0} \frac{16\log(n)}{\Delta_i}.$$

**Algorithm 1** UCB($\delta$)

**Input:** $n \in \mathbb{N}$, $\delta > 0$

**for** $t = 1, \ldots, n$ **do**

Choose action $A_t = \arg\max_i UCB_i(t-1,\delta)$.

Observe reward $X_t$ and update upper confidence bounds.

**end for**

In the proof we will use the following notations: If $(X_{ti})_{t\in[n],i\in[k]}$ is a collection of independent random variables with distribution $\mathbb{P}_i$, then define $\hat{\mu}_{is} = \frac{1}{s}\sum_{u=1}^{s} X_{ui}$ to be the sample mean based on the first $s$ samples. The arm chosen in round $t$ will be denoted by $A_t$.

*Proof.* We may assume, without the loss of generality, that the first arm is optimal, so $\mu_1 = \mu^*$. By the regret decomposition lemma,

$$R_n = \sum_{i=1}^{k} \Delta_i \mathbb{E}\left[T_i(n)\right]. \tag{1}$$

We will prove the theorem by bounding $\mathbb{E}\left[T_i(n)\right]$ for each suboptimal arm. We decouple the randomness from the behavior of the UCB algorithm. Let $G_i$ be the event defined by

$$G_i = \left\{\mu_1 < \min_{t\in[n]} UCB_1(t,\delta)\right\} \cap \left\{\hat{\mu}_{iu_i} + \sqrt{\frac{2}{u_i}\log\left(\frac{1}{\delta}\right)} < \mu_1\right\},$$

where $u_i \in [n]$ is a constant whose value will be determined later. So $G_i$ is the event when $\mu_1$ is not underestimated in any of the rounds by the upper confidence bound of the first arm, while the upper confidence bound for the mean of arm $i$ after taking $u_i$ samples from arm $i$ is below the expectation of the optimal arm. We will show the following two points:

- If $G_i$ occurs, the arm $i$ will be played at most $u_i$ times: $T_i(n) \le u_i$.
- The complement event $G_i^c$ occurs with low probability.

Since $T_i(n) \le n$,

$$\mathbb{E}[T_i(n)] = \mathbb{E}[\mathbb{I}\{G_i\}T_i(n)] + \mathbb{E}[\mathbb{I}\{G_i^c\}T_i(n)] \le u_i + \mathbb{P}(G_i^c)\, n. \tag{2}$$

Now we will show that $T_i(n) \le u_i$ on $G_i$. We prove this by contradiction. Suppose that $T_i(n) > u_i$. In this case arm $i$ was played more than $u_i$ times over the $n$ rounds,

so there exists a round $t \in [n]$ where $T_i(t-1) = u_i$ and $A_t = i$. By definition of $G_i$,

$$\begin{aligned} UCB_i(t-1,\delta) &= \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}} \\ &= \hat{\mu}_{iu_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}} \\ &< \mu_1 \\ &< UCB_1(t-1,\delta). \end{aligned}$$

This implies that $A_t = \arg\max_j UCB_j(t-1,\delta) \neq i$, which is a contradiction. This proves that if $G_i$ occurs, then $T_i(n) \leq u_i$. Now we will prove that $G_i^c$ occurs with low probability. By the definition of $G_i^c$,

$$G_i^c = \left\{\mu_1 \geq \min_{t\in[n]} UCB_1(t,\delta)\right\} \cup \left\{\hat{\mu}_{iu_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}} \geq \mu_1\right\}. \tag{3}$$

We decompose the first set using the definition of $UCB_1(t,\delta)$,

$$\begin{aligned} \left\{\mu_1 \geq \min_{t\in[n]} UCB_1(t,\delta)\right\} \subset &\left\{\mu_1 \geq \min_{s\in[n]} \hat{\mu}_{1s} + \sqrt{\frac{2\log(1/\delta)}{s}}\right\} \\ &\bigcup_{s\in[n]} \left\{\mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2\log(1/\delta)}{s}}\right\} \end{aligned}$$

Using Theorem 2 to bound the probability that the difference between the expectation and the sample mean is large, we obtain:

$$\begin{aligned} \mathbb{P}\left(\mu_1 \geq \min_{t\in[n]} UCB_1(t,\delta)\right) &\leq \mathbb{P}\left(\bigcup_{s\in[n]} \left\{\mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2\log(1/\delta)}{s}}\right\}\right) \\ &\leq \sum_{s=1}^{n} \mathbb{P}\left(\mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2\log(1/\delta)}{s}}\right) \\ &\leq n\delta. \end{aligned} \tag{4}$$

The next step is to bound the probability of this set

$$\left\{\hat{\mu}_{iu_i} + \sqrt{\frac{2}{u_i}\log\left(\frac{1}{\delta}\right)} \geq \mu_1\right\}.$$

Assume that $u_i$ is large enough that

$$\Delta_i - \sqrt{\frac{2\log(1\delta)}{u_i}} \geq c\Delta_i \tag{5}$$

for some constant $c \in (0,1)$ whose value will be determined later. Since $\mu_1 = \mu_i + \Delta_i$, using Theorem 2 we get that

$$\begin{aligned}
\mathbb{P}\left(\hat{\mu}_{iu_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}} \geq \mu_1\right) &= \mathbb{P}\left(\hat{\mu}_{iu_i} - \mu_i \geq \Delta_i - \sqrt{\frac{2\log(1/\delta)}{u_i}}\right) \\
&\leq \mathbb{P}(\hat{\mu}_{iu_i} - \mu_i \geq c\Delta_i) \\
&\leq \exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right).
\end{aligned}$$

Combining this with (3) and (4) gives that

$$\mathbb{P}(G_i^c) \leq n\delta + \exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right).$$

If we substitute this into (2) we obtain the following result:

$$\mathbb{E}[T_i(n)] \leq u_i + n\left(n\delta + \exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right)\right). \tag{6}$$

We have to choose $u_i \in [n]$ to satisfy (5). It is natural to choose the smallest integer for which (5) holds, which is

$$u_i = \left\lceil \frac{2\log(1/\delta)}{(1-c)^2\Delta_i^2} \right\rceil.$$

If $u_i > n$, then (6) holds trivially, since $T_i(n) \leq n$. Using the assumption that $\delta = 1/n^2$ and the choice of $u_i$ we get that

$$\mathbb{E}[T_i(n)] \leq u_i + 1 + n^{1-2c^2/(1-c)^2} = \left\lceil \frac{2\log(n^2)}{(1-c)^2\Delta_i^2} \right\rceil + 1 + n^{1-2c^2/(1-c^2)}. \tag{7}$$

It remains to choose $c \in (0,1)$. Unless $2c^2/(1-c^2) \geq 1$, the second term will contribute a polynomial dependence on $n$. If $c$ is too close to 1, the first term will be large. Choosing $c = 1/2$ leads to

$$\mathbb{E}[T_i(n)] \leq 3 + \frac{16\log(n)}{\Delta_i^2}.$$

Substituting this into (1) we get that

$$R_n = \sum_{i=1}^{k} \Delta_i \mathbb{E}\left[T_i(n)\right] \leq 3 \sum_{i=1}^{k} \Delta_i + \sum_{i:\Delta_i > 0} \frac{16 \log(n)}{\Delta_i}. \tag{8}$$

□

The bound on the regret in Theorem 3 depends on the reciprocal of the gaps, and if there is a very small suboptimality gap, then the bound will be large. However, we can also prove a sublinear regret bound, that does not depend on the reciprocal of the gaps.

**Theorem 4.** *If $\delta = 1/n^2$, then the regret of UCB on any 1-subgaussian environment with $k$ arms, is bounded by*

$$R_n \leq 8\sqrt{nk \log(n)} + 3 \sum_{i=1}^{k} \Delta_i.$$

*Proof.* Let $\Delta > 0$, its value will be determined later. From the proof of Theorem 3 we know that for each suboptimal arm $i$,

$$\mathbb{E}[T_i(n)] \leq 3 + \frac{16 \log(n)}{\Delta_i^2}.$$

Using the regret decomposition lemma (Lemma 1), we get that

$$\begin{aligned}
R_n &= \sum_{i=1}^{k} \Delta_i \mathbb{E}\left[T_i(n)\right] \\
&= \sum_{i:\Delta_i < \Delta} \Delta_i \mathbb{E}\left[T_i(n)\right] + \sum_{i:\Delta_i \geq \Delta} \Delta_i \mathbb{E}\left[T_i(n)\right] \\
&\leq n\Delta + \sum_{i:\Delta_i \geq \Delta} \left(3\Delta_i + \frac{16 \log(n)}{\Delta_i}\right) \\
&\leq n\Delta + \frac{16k \log(n)}{\Delta} + 3 \sum_{i} \Delta_i \\
&\leq 8\sqrt{nk \log(n)} + 3 \sum_{i=1}^{k} \Delta_i.
\end{aligned}$$

Here the first inequality is true, because $\sum_{i:\Delta_i < \Delta} T_i(n) \leq n$ and the last inequality follows by choosing $\Delta = \sqrt{nk \log(n)/n}$. □

The additive $\sum_i \Delta_i$ term is unavoidable because the algorithm has to play each arm at least once. This term does not grow as $n$ grows, and it is typically negligible.

## 3.4 Asymptotic Optimality

The UCB Algorithm we defined previously have to run for a fixed number of steps to return an arm. It can be modified to be an anytime algorithm, meaning that the algorithm returns a valid solution to a problem even if it is interrupted before it ends. The algorithm is expected to find a better solution if it keep running longer.

By changing the UCB Algorithm we defined earlier, we can achieve a bound, which has the same order as in Theorem 3 section but the leading constant, governing the asymptotic rate of growth of the regret is smaller. In Algorithm 2 number of arms is denoted by $k$ and $f(t) = 1 + t\log^2(t)$.

**Theorem 5.** *For any 1-subgaussian bandit, the regret of Algorithm 2 satisfies*

$$R_n \leq \sum_{i:\Delta_i>0} \inf_{\varepsilon\in(0,\Delta_i)} \Delta_i \left(1 + \frac{5}{\varepsilon^2} + \frac{2\left(\log f(n) + \sqrt{\pi \log f(n)} + 1\right)}{(\Delta_i - \varepsilon)^2}\right).$$

From this inequality we can see that

$$\limsup_{n\to\infty} \frac{R_n}{\log(n)} \leq \sum_{i:\Delta_i>0} \frac{2}{\Delta_i}.$$

A lower bound can be given, which is a constant multiple of the above expression and this proves the asymptotic optimality of the algorithm.

**Algorithm 2** UCB($\delta$)

**Input:** $k \in \mathbb{N}$

Choose each arm once.

Choose action $A_t = \arg\max_i \left(\hat{\mu}(t-1) + \sqrt{\frac{2\log f(t)}{T_i(t-1)}}\right)$.

# 4 Median Elimination

In this section we will deal with a new learning objective, when the goal is to develop an algorithm that finds a near optimal arm with high probability. In this case the horizon is not fixed, and $n$ will be used to denote the number of arms. An arm is called optimal, if it has the highest expected reward among all of the arms. There can be multiple optimal arms, their expected reward is denoted by $r^*$.

**Definition 2.** *An arm $a \in \mathcal{A}$ is called $\varepsilon$-optimal if*

$$\mathbb{E}[X(a)] \geq r^* - \varepsilon,$$

*where $X(a)$ is a reward sampled from the distribution of arm $a$.*

**Definition 3.** *An algorithm is called an $(\varepsilon, \delta)$-PAC (probably approximately correct) algorithm for the multi-armed bandit problem with sample complexity $T$, if it outputs an $\varepsilon$-optimal arm with probability at least $1 - \delta$ when it terminates, and the number of steps the algorithm performs until termination is bounded by $T$.*

## 4.1 Algorithm

We state a version of the Median Elimination algorithm [13, 14], for the case, when the arms are 1-subgaussian.

**Algorithm 3** Median Elimination

**Input:** $\varepsilon > 0$, $\delta > 0$

**Output:** an arm which is $\varepsilon$-optimal with probability at least $1 - \delta$

Set $S_1 = \mathcal{A}$, $\varepsilon_1 = \varepsilon/4$, $\delta_1 = \delta/2$, $\ell = 1$.

**repeat**

Sample every arm $a \in S_\ell$ for $n_\ell = \lceil 8 \log(3/\delta_\ell)/\varepsilon_\ell^2 \rceil$ times, and let $\hat{\mu}_a^\ell$ denote its sample mean.

Find the median of the $\hat{\mu}_a^\ell$ values and denote it by $m_\ell$.

$S_{\ell+1} = S_\ell \setminus \{a : \hat{\mu}_a^\ell < m_\ell\}$

$\varepsilon_{\ell+1} = \frac{3}{4}\varepsilon_\ell$, $\delta_{\ell+1} = \delta_\ell/2$, $\ell = \ell + 1$

**until** $|S_\ell| = 1$

## 4.2 Sample Complexity of Median Elimination

**Theorem 6.** *The Median Elimination algorithm is an $(\varepsilon, \delta)$-PAC algorithm and its sample complexity is*

$$\mathcal{O}\left(\frac{n}{\varepsilon^2}\log\frac{1}{\delta}\right).$$

We will use two lemmas to prove this theorem.

**Lemma 2.** *In the Median Elimination algorithm, for every phase $\ell$*

$$\mathbb{P}\left(\max_{j\in S_\ell}\mu_j \le \max_{i\in S_{\ell+1}}\mu_i + \varepsilon_\ell\right) \ge 1-\delta_\ell.$$

*Proof.* Without the loss of generality we can look at the first round and assume that $\mu_1$ is the expectation of the best arm. We bound the probability of failure by looking at the event $E_1 = \{\hat{\mu}_1 < \mu_1 - \varepsilon_1/2\}$, when the sample mean underestimates the expectation. We sample the arms so many times in the first round, that

$$\begin{aligned}
\mathbb{P}(E_1) &= \mathbb{P}\left(\hat{\mu}_1 < \mu_1 - \varepsilon_1/2\right) \\
&\le \exp\left(-\frac{n_1}{2}\left(\frac{\varepsilon_1}{2}\right)^2\right) \\
&= \exp\left(-\frac{1}{2}\left\lceil\frac{8\log(3/\delta_1)}{\varepsilon_1^2}\right\rceil\left(\frac{\varepsilon_1}{2}\right)^2\right) \\
&\le \exp\left(-\frac{1}{2}\frac{8\log(3/\delta_1)}{\varepsilon_1^2}\left(\frac{\varepsilon_1}{2}\right)^2\right) \\
&= \frac{\delta_1}{3}.
\end{aligned}$$

In the first inequality we used the fact, that the arms are 1-subgaussian, then we used the definition of $n_\ell$. Now we bound the probability, that the sample mean of arm $j$ is larger than the sample mean of the optimal arm, if $E_1$ does not hold and arm $j$ is not $\varepsilon_1$-optimal. If $\hat{\mu}_1 \ge \mu_1 - \varepsilon_1/2$, $\mu_j \le \mu_1 - \varepsilon_1$ and $\hat{\mu}_j \ge \hat{\mu}_1$, then

$$\hat{\mu}_j \ge \hat{\mu}_1 \ge \mu_1 - \varepsilon_1/2 \ge \mu_j + \varepsilon_1/2.$$

this way

$$\mathbb{P}\left(\hat{\mu}_j \ge \hat{\mu}_1 \,|\, \hat{\mu}_1 \ge \mu_1 - \varepsilon_1/2\right) \le \mathbb{P}\left(\hat{\mu}_j \ge \mu_j + \varepsilon_1/2 \,|\, \hat{\mu}_1 \ge \mu_1 - \varepsilon_1/2\right) \le \delta_1/3. \tag{9}$$

Let $Y$ denote the number of arms which are not $\varepsilon_1$-optimal but have a higher sample mean than the optimal arm. Based on (9), $\mathbb{E}[Y \mid \hat{\mu}_1 \geq \hat{\mu}_1 - \varepsilon_1/2] \leq n\delta_1/3$. By applying Markov's inequality, we get that

$$\mathbb{P}(Y \geq n/2 \mid \hat{\mu}_1 \geq \hat{\mu}_1 - \varepsilon_1/2) \leq \frac{n\delta_1/3}{n/2} = 2\delta_1/3.$$

This way the probability of failure is bounded by $\delta_1$. □

**Lemma 3.** *The sample complexity of Median Elimination is*

$$\mathcal{O}\left(\frac{n}{\varepsilon^2}\log\frac{1}{\delta}\right).$$

*Proof.* In the $\ell$-th round $\frac{n}{2^{\ell-1}}$ arms are sampled, so the number of arms sampled in total is $\frac{n}{2^{\ell-1}}\frac{8\log(3/\delta_\ell)}{\varepsilon_\ell^2}$. By definition $\delta_\ell = \delta/2^\ell$ and $\varepsilon_\ell = \left((3/4)^{\ell-1}\varepsilon/4\right)$. We can bound the sample complexity the following way:

$$\begin{aligned}
\sum_{\ell=1}^{\log_2(n)} \frac{n}{2^{\ell-1}}\frac{8\log(3/\delta_\ell)}{\varepsilon_\ell^2} &= 8\sum_{\ell=1}^{\log_2(n)} \frac{n}{2^{\ell-1}}\frac{\log(2^\ell 3/\delta)}{\left((3/4)^{\ell-1}\varepsilon/4\right)^2} \\
&= \frac{128n}{\varepsilon^2}\sum_{\ell=1}^{\log_2(n)} \left(\frac{8}{9}\right)^{\ell-1}\left(\log\frac{1}{\delta} + \log 3 + \ell\log 2\right) \\
&\leq \frac{128n\log(1/\delta)}{\varepsilon^2}\sum_{\ell=1}^{\log_2(n)} \left(\frac{8}{9}\right)^{\ell-1}(\ell C' + C) \\
&= \mathcal{O}\left(\frac{n}{\varepsilon^2}\log\frac{1}{\delta}\right).
\end{aligned}$$

□

*Proof of Theorem 6.* By Lemma 3, the sample complexity is bounded by $\mathcal{O}\left(\frac{n}{\varepsilon^2}\log\frac{1}{\delta}\right)$. By Lemma 2, the probability of failure in round $i$ is $\delta_i$, so the probability of failure over all rounds is bounded by $\sum_{\ell=1}^{\log_2(n)}\delta_i \leq \delta$. In each round, the optimal reward of the remaining arms is reduced by $\varepsilon_i$ at most, so the total error is bounded by $\sum_{\ell=1}^{\log_2(n)}\varepsilon_i \leq \varepsilon$. □

## 4.3 Lower Bound on the Sample Complexity

As we saw in the previous section, $\mathcal{O}\left((n/\varepsilon^2)\log(1/\delta)\right)$ sampling is enough to find an $\varepsilon$-optimal arm with probability at least $1-\delta$ if there are $n$ arms. In this section we will provide a matching lower bound on the expected number of trials under any $(\varepsilon,\delta)$-PAC policy. The results of this section can be found in [15].

In this section, we will examine the special case when the rewards are Bernoulli random variables with unknown parameters (biases). In this special case the arms are often called coins and the unit rewards will be called heads. The number of arms will be denoted by $n$, and $X_k^\ell$ will denote the reward obtained the $k$th time arm $\ell$ is sampled, $p_\ell$ will denote the mean of arm $\ell$ and $p=(p_1,...,p_n)$. We will only consider policies that stop with probability 1 for every possible $p$. Given a policy, $\mathbb{P}_p$ will denote the corresponding probability measure on the natural probability space for this model, which captures both the randomness in the arms and the randomization in the policy. The number of times arm $\ell$ is sampled will be denoted by $T_\ell$ and $T$ will denote the total number of trials.

**Theorem 7.** *There exist positive constants $c_1$, $c_2$, $\varepsilon_0$ and $\delta_0$ such that for every $n\geq 2$, $\varepsilon\in(0,\varepsilon_0)$, $\delta\in(0,\delta_0)$ and for every $(\varepsilon,\delta)$-PAC policy, there exists a $p\in[0,1]^n$ such that*

$$\mathbb{E}_p[T]\geq c_1\frac{n}{\varepsilon^2}\log\frac{c_2}{\delta}.$$

*In particular, $\varepsilon_0=1/8$ and $\delta_0=e^{-4}/4$ satisfies this.*

*Proof.* We define $\varepsilon_0=1/8$ and $\delta_0=\varepsilon^{-4}/4$ and fix some $\varepsilon\in(0,\varepsilon_0)$, $\delta\in(0,\delta_0)$. Let us consider the multi-armed bandit problem with $n+1$ coins numbered from 0 to $n$. We will consider a finite set of possible $p$ vectors, and we will call these hypotheses. The bias of coin 0 will be $p_0=(1+\varepsilon)/2$ under all of the hypotheses. Under $H_0$, all the other coins have a bias of $1/2$,

$$H_0: p_0=\frac{1}{2}+\frac{\varepsilon}{2},\quad p_i=\frac{1}{2},\ \forall i\neq 0.$$

For $\ell=1,...,n$, define

$$H_\ell: p_0=\frac{1}{2}+\frac{\varepsilon}{2},\quad p_\ell=\frac{1}{2}+\varepsilon,\quad p_i=\frac{1}{2},\ \forall i\neq 0,\ell.$$

Fix an $(\varepsilon/2, \delta)$-PAC policy. Under $H_\ell$, the policy must select coin $\ell$ with probability at least $1-\delta$. Under $H_\ell$, the expectation and the probability will be denoted by $\mathbb{E}_\ell$ and $\mathbb{P}_\ell$. We define

$$t^* = \frac{1}{c\varepsilon^2}\log\frac{1}{4\delta} = \frac{1}{c\varepsilon^2}\log\frac{1}{\theta},$$

where $c$ is a constant whose value will be chosen later and $\theta = 4\delta$. This way $\theta < e^{-4}$. Let us assume, that there exist a coin $\ell \neq 0$, for which $\mathbb{E}_0[T_\ell] \leq t^*$. We will show, that in this case, the probability of selecting $H_0$ under $H_\ell$ is greater than $\delta$ and it contradicts the fact that the policy is $(\varepsilon/2, \delta)$-PAC. Without the loss of generality we may assume, that $\mathbb{E}_0[T_1] \leq t^*$. Let us define $A = \{T_1 \leq 4t^*\}$.
Since $t^* \geq \mathbb{E}_0[T_1] \geq 4t^*\mathbb{P}_0(T_1 > 4t^*) = 4t^*(1 - P_0(T_1 \leq 4t^*))$, we obtain that $P_0(A) \geq 3/4$. We define $K_t = X_1^1 + ... + X_t^1$ to be the number of heads if coin 1 is sampled $t$ times. We define event $C$ the following way:

$$C = \left\{ \max_{1\leq t\leq 4t^*} \left| K_t - \frac{1}{2}t \right| < \sqrt{t^*\log(1/\theta)} \right\}.$$

Now we will prove a few lemmas which will be used for the proof of this theorem.

**Lemma 4.** $\mathbb{P}_0(C) > \frac{3}{4}$.

*Proof.* We will prove a more general result. Let us denote the bias of coin $i$ under hypothesis $H_\ell$ by $p_i$, and the number of heads if coin $i$ is tossed $t$ times by $K_t^i$. Let us define $C_i$ the following way:

$$C_i = \left\{ \max_{1\leq t\leq 4t^*} \left| K_t^i - p_i t \right| < \sqrt{t^*\log(1/\theta)} \right\}.$$

It is easy to check that $K_t^i - p_i t$ is a $\mathbb{P}_\ell$-martingale. Since $x \to x^2$ is a convex function, and a martingale composed with a convex function is a submartingale, we can use Doob's maximal inequality to bound the probability of the complement of $C_i$:

$$\begin{aligned}\mathbb{P}_\ell\left( \max_{1\leq t\leq 4t^*} \left| K_t^i - p_i t \right| \geq \sqrt{t^*\log(1/\theta)} \right) &= \mathbb{P}_\ell\left( \max_{1\leq t\leq 4t^*} \left( K_t^i - p_i t \right)^2 \geq t^*\log(1/\theta) \right) \\ &\leq \frac{\mathbb{E}_\ell[(K_{4t^*}^i - 4p_i t^*)^2]}{t^*\log(1/\theta)}.\end{aligned}$$

Since $\mathbb{E}_\ell[(K_{4t^*}^i - 4p_i t^*)^2] = 4p_i(1-p_i)t^*$, we get that

$$\mathbb{P}_\ell(C_i) \geq 1 - \frac{4p_i(1-p_i)}{\log(1/\theta)} > \frac{3}{4}.$$

The last inequality is true, because $\theta < e^{-4}$ and $4p_i(1-p_i) \leq 1$.

Under hypothesis $H_0$, coin $i = 1$ has bias $p_i = 1/2$. Applying the result under these settings we get that $\mathbb{P}_0(C) > 3/4$. □

**Lemma 5.** *If $0 \leq x \leq 1/\sqrt{2}$ and $y \geq 0$, then $(1-x)^y \geq e^{-dxy}$, where $d = 1.78$.*

*Proof.* It is straightforward to prove that, $\log(1-x) + dx \geq 0$, for $0 \leq x \leq 1/\sqrt{2}$. This way, $y\left(\log(1-x) + dx\right) \geq 0$ for all $y \geq 0$. By rearranging and exponentiating, we get that $(1-x)^y \geq e^{-dxy}$. □

Let $B$ be the event, that the policy selects coin 0. Since the policy is $(\varepsilon/2, \delta)$-PAC and $\delta < e^{-4}/4 < 1/4$, the probability of $B$ is $\mathbb{P}_0(B) > 3/4$. We have already seen that $P_0(A) \geq 3/4$ and $P_0(C) > 3/4$. Let us define $S$ to be the intersection of $A$, $B$ and $C$, so $S = A \cap B \cap C$. Based on the above $P_0(S) > 1/4$.

**Lemma 6.** *If $\mathbb{E}_0[T_1] \leq t^*$ and $c \geq 100$, then $\mathbb{P}_1(B) > \delta$.*

*Proof.* Let $W$ denote the history of the process until the policy terminates (the sequence of selected coins at each time step and the observed rewards). We denote the likelihood function by $L_\ell(w) = \mathbb{P}_\ell(W = w)$. We will use $K$ to denote the total number of heads obtained from coin 1. If the history is given up to round $t-1$, the coin choice at time $t$ has the same probability under $H_0$ and $H_1$ and the reward at round $t$ has the same distribution under $H_0$ and $H_1$, unless the chosen coin in round $t$ is coin 1. This way,

$$\begin{aligned} \frac{L_1(W)}{L_0(W)} &= \frac{(1/2+\varepsilon)^K (1/2-\varepsilon)^{T_1-K}}{(1/2)^{T_1}} \\ &= (1+2\varepsilon)^K (1-2\varepsilon)^K (1-2\varepsilon)^{T_1-2K} \\ &= (1-4\varepsilon^2)^K (1-2\varepsilon)^{T_1-2K}. \end{aligned} \tag{10}$$

Now we will give a lower bound of this expression, when event $S$ occurs. In this case event $A$ occurs as well, so $K \leq T_1 \leq 4t^*$ and

$$\begin{aligned} (1-4\varepsilon^2)^K &\geq (1-4\varepsilon^2)^{4t^*} \\ &= (1-4\varepsilon^2)^{4\log(1/\theta)/(c\varepsilon)^2} \\ &\geq e^{-16d\log(1/\theta)/c} \\ &= \theta^{16d/c}. \end{aligned}$$

In the second inequality we used Lemma 5, which applies, since $4\varepsilon^2 < 4/4^2 < 1/\sqrt{2}$. If event $S$ occurs, then event $A \cap C$ occurs as well, so using the definition of $t^*$ we get that

$$T_1 - 2K \leq 2\sqrt{t^* \log(1/\theta)} = (2/\varepsilon\sqrt{c}) \log(1/\theta).$$

Using this, we can give a lower bound on $(1-2\varepsilon)^{T_1-K}$:

$$\begin{aligned}(1-2\varepsilon)^{T_1-K} &\geq (1-2\varepsilon)^{(2/\varepsilon\sqrt{c})\log(1/\theta)} \\ &\geq e^{-(4d/\sqrt{c})\log(1/\theta)} \\ &= \theta^{4d/\sqrt{c}}.\end{aligned}$$

If we substitute this into (10) we get that

$$\frac{L_1(W)}{L_0(W)} \geq \theta^{16d/c+4d/\sqrt{c}}.$$

If we pick a large enough $c$, for example $c = 100$, we get that

$$L_1(W)/L_0(W) > \theta = 4\delta,$$

whenever $S$ occurs. This way

$$\frac{L_1(W)}{L_0(W)}\mathbb{I}_S \geq 4\delta\mathbb{I}_S,$$

where $\mathbb{I}_S$ denotes the indicator of event $S$. Combining this result with the fact that $\mathbb{P}_0(S) > 1/4$, we get that

$$\mathbb{P}_1(B) \geq \mathbb{P}_1(S) = \mathbb{E}[\mathbb{I}_S] = \mathbb{E}_0\left[\frac{L_1(W)}{L_0(W)}\mathbb{I}_S\right] \geq \mathbb{E}_0[4\delta\mathbb{I}_S] = 4\delta\mathbb{P}_0(S) > \delta.$$

□

We have shown that $\mathbb{P}_1(B) > \delta$ if $c \geq 100$ and

$$\mathbb{E}_0[T_1] \leq \frac{1}{c\varepsilon^2}\log\frac{1}{4\delta}.$$

So if we have an $(\varepsilon/2, \delta)$-PAC policy, then for every $\ell > 0$

$$\mathbb{E}_0[T_\ell] > \frac{1}{c\varepsilon^2}\log\frac{1}{4\delta}.$$

Equivalently, for every $(\varepsilon, \delta)$-PAC policy

$$\mathbb{E}_0[T] > \frac{1}{4c\varepsilon^2}\log\frac{1}{4\delta}.$$

□

# 5 Finitely Many Arms with a Special Structure

In this section we will consider a special case of the multi armed bandit problem under the PAC settings in which there are finitely many arms and we have further knowledge about the structure of the arms. This special case arises in applications, for example the quantized estimation problem studied in [16] leads to a bandit problem of this kind. In this case the assumptions on the arms are the following:

**Assumption 1.** *There are* $n = 2^m + 1$, $m \geq 1$ *arms.*

The arms will be denoted by $a_0, a_1, ..., a_{2^m}$ and the expectation of arm $a_i$ will be denoted by $\mu_i$.

**Assumption 2.** *There exists a* $k \in \{0, 1, ..., 2^m\}$ *such that*

$$\mu_0 < \mu_1 < ... < \mu_{k-1} < \mu_k > \mu_{k+1} > \mu_{k+2} > ... > \mu_{2^m}.$$

**Assumption 3.** *There exists a known* $\Delta > 0$*, such that for all* $0 \leq i \leq 2^m - 1$,

$$|\mu_{i+1} - \mu_i| \geq \Delta.$$

**Assumption 4.** *The arms are* 1*-subgaussian.*

## 5.1 Algorithm and Sample Complexity

For a fixed $\delta \in (0, 1)$, we can find the optimal arm with probability at least $1 - \delta$ by running Median Elimination with $\varepsilon = \Delta/2$. The sample complexity of this solution is

$$\mathcal{O}\left(\frac{n}{\Delta^2} \log \frac{1}{\delta}\right).$$

However under these assumption, we can create an algorithm, which finds the optimal arm much faster.

**Algorithm 4**

**Input:** $\delta > 0$

**Output:** an arm which is optimal with probability at least $1 - \delta$

Set $S_1 = \mathcal{A}$, $\delta_1 = \delta/2$, $\ell = 1$.

**while** $|S_\ell| > 3$ **do**

Sample $n_\ell = \lceil \log(4/\delta_\ell)/(2^{2m-2\ell-5}\Delta^2) \rceil$ times each of the three arms:

$a_j$, $j \in \{i \cdot 2^{m-\ell-1},\, i = 1, 2, 3\}$ and let $\hat{\mu}_j^\ell$ denote their sample means.

$i_\ell^* = \arg\max_j \hat{\mu}_j^\ell$

$S_{\ell+1} = \left\{a_i : i_\ell^* - 2^{m-\ell-1} \leq i \leq i_\ell^* + 2^{m-\ell-1}\right\}$

Renumber the arms from 0 to $2^{m-\ell}$.

$\delta_{\ell+1} = \delta_\ell/2$, $\ell = \ell + 1$

**end while**

Sample $n_m = \lceil \log(4/\delta_m)/(2^{-3}\Delta^2) \rceil$ times the remaining arms: $a_0, a_1, a_2$.

Let $\hat{\mu}_j^m$ denote their sample means.

$i_m^* = \arg\max_{j \in \{0,1,2\}} \hat{\mu}_j^m$

**return** $a_{i_m^*}$

**Theorem 8.** *Under Assumptions 1-4, Algorithm 4 finds the optimal arm with probability at least* $1 - \delta$ *and its sample complexity is*

$$\mathcal{O}\left(\log n + \frac{1}{\Delta^2} \log \frac{n}{\delta}\right).$$

We will use the following lemmas to prove Theorem 8.

**Lemma 7.** *For every phase* $\ell = 1, 2, ..., m-1$

$$\mathbb{P}\left(\max_{j \in S_\ell} \mu_j > \max_{i \in S_{\ell+1}} \mu_i\right) \leq \delta_\ell.$$

*Proof.* Let $i^* = \arg\max_{j \in S_\ell} \mu_j$. We have to show that $\mathbb{P}\left(a_{i^*} \notin S_{\ell+1}\right) \leq \delta_\ell$.
Since $a_{i^*} \notin S_{\ell+1}$ if and only if $|i^* - i_\ell^*| > 2^{m-\ell-1}$,

$$\begin{aligned}\mathbb{P}\left(a_{i^*} \notin S_{\ell+1}\right) &= \mathbb{P}\left(|i^* - i_\ell^*| > 2^{m-\ell-1}\right) \\ &= \sum_{i=1}^{3} \left[\mathbb{P}(|i^* - i_\ell^*| > 2^{m-\ell-1} \,|\, i_\ell^* = i \cdot 2^{m-\ell-1}) \cdot \mathbb{P}\left(i_\ell^* = i \cdot 2^{m-\ell-1}\right)\right].\end{aligned}$$

The value of $\mathbb{P}\left(|i^* - i^*_\ell| > 2^{m-\ell-1} \,|\, i^*_\ell = i \cdot 2^{m-\ell-1}\right)$ is either 0 or 1 and it is 0 for at least one value of $i \in \{1, 2, 3\}$. We have to give an upper bound on the probability $\mathbb{P}\left(i^*_\ell = i \cdot 2^{m-\ell-1}\right)$ when $|i^* - i \cdot 2^{m-\ell-1}| > 2^{m-\ell-1}$. We will show that in this case $\mathbb{P}\left(i^*_\ell = i \cdot 2^{m-\ell-1}\right) \leq \delta_\ell/2$ and from this the statement of the lemma follows.

First we deal with the case, when $i^* > i \cdot 2^{m-\ell-1}$. With the $j = (i+1) \cdot 2^{m-\ell-1}$, $i' = i \cdot 2^{m-\ell-1}$ notations: $i^* > j > i' \implies \mu_{i^*} > \mu_j > \mu_{i'}$, because $i^*$ is the index of the optimal arm in $S_\ell$ and based on Assumption 2 the arms satisfy that

$$\mu_0 < \mu_1 < ... < \mu_{i^*} > ... > \mu_{2^{m+1-\ell}}.$$

Since $j - i' = 2^{m-\ell-1}$, from Assumption 2 and 3 follows that $\mu_j \geq \mu_{i'} + 2^{m-\ell-1}\Delta$. By the definition of $i^*_\ell$ if $i^*_\ell = i'$ then $\hat{\mu}^\ell_{i'} \geq \hat{\mu}^\ell_j$, so $\mathbb{P}(i^*_\ell = i') \leq \mathbb{P}(\hat{\mu}^\ell_{i'} \geq \hat{\mu}^\ell_j)$. Consider the following events:

$$\begin{aligned} A &= \{\hat{\mu}^\ell_j > \mu_j - 2^{m-\ell-2}\Delta\}, \\ B &= \{\hat{\mu}^\ell_{i'} < \mu_{i'} + 2^{m-\ell-2}\Delta\}, \\ C &= \{\hat{\mu}^\ell_j > \hat{\mu}^\ell_{i'}\}. \end{aligned}$$

It is easy to see that $A \wedge B \implies C$ :

$$\hat{\mu}^\ell_j > \mu_j - 2^{m-\ell-2}\Delta \geq \mu_{i'} + 2^{m-\ell-1}\Delta - 2^{m-\ell-2}\Delta = \mu_{i'} + 2^{m-\ell-2}\Delta > \hat{\mu}^\ell_{i'}.$$

This implies that:

$$\mathbb{P}(i^*_\ell = i') \leq \mathbb{P}(\hat{\mu}^\ell_{i'} \geq \hat{\mu}^\ell_j) = \mathbb{P}(\overline{C}) \leq \mathbb{P}(\overline{A} \vee \overline{B}) \leq \mathbb{P}(\overline{A}) + \mathbb{P}(\overline{B}).$$

It remains to show that $\mathbb{P}(\overline{A}) \leq \delta_\ell/4$ and $\mathbb{P}(\overline{B}) \leq \delta_\ell/4$:

$$\begin{aligned} \mathbb{P}(\overline{A}) &= \mathbb{P}(\hat{\mu}^\ell_j \leq \mu_j - 2^{m-\ell-2}\Delta) \\ &\leq \exp\left(-\frac{1}{2}(2^{m-\ell-2}\Delta)^2 \left\lceil \frac{\log(4/\delta_\ell)}{2^{2m-2\ell-5}\Delta^2} \right\rceil\right) \\ &\leq \exp\left(-\frac{1}{2}(2^{m-\ell-2}\Delta)^2 \frac{\log(4/\delta_\ell)}{2^{2m-2\ell-5}\Delta^2}\right) \\ &= \delta_\ell/4. \end{aligned}$$

Similarly, $\mathbb{P}(\overline{B}) \leq \delta_\ell/4$. The case when $i^* < i \cdot 2^{m-\ell-1}$ can be proved similarly. $\square$

**Lemma 8.** *For the final phase:*

$$\mathbb{P}(\max_{j \in S_m} \mu_j > \mu_{i^*_m}) \leq \delta_m.$$

*Proof.* Let $i^* = \arg\max_{j \in S_m} \mu_j$. We have to show that $\mathbb{P}\left(i^* \neq i^*_m\right) \leq \delta_m$. By the law of total probability,

$$\mathbb{P}(i^* \neq i^*_m) = \sum_{j=0}^{2} \mathbb{P}(i^* \neq i^*_m \,|\, i^*_m = j)\, \mathbb{P}(i^*_m = j).$$

The value of $\mathbb{P}(i^* \neq i^*_m \,|\, i^*_m = j)$ is either 0 or 1 and it is 1 for exactly two values of $j$. We prove that $\mathbb{P}(i^*_m = j) \leq \delta_m/2$ when $j \neq i^*$ and this proves the lemma. By the definition of $i^*_m$, if $i^*_m = j$ then $\hat{\mu}^m_j \geq \hat{\mu}^m_{i^*}$. This way:

$$\mathbb{P}(i^*_m = j) \leq \mathbb{P}(\hat{\mu}^m_j \geq \hat{\mu}^m_{i^*}).$$

By the definition of $i^*$ : $\mu_j \leq \mu_{i^*} - \Delta$. Consider the following events:

$$\begin{aligned} A &= \{\hat{\mu}^m_{i^*} > \mu_{i^*} - \Delta/2\}, \\ B &= \{\hat{\mu}^m_j < \mu_j + \Delta/2\}, \\ C &= \{\hat{\mu}^m_{i^*} > \hat{\mu}^m_j\}. \end{aligned}$$

It is easy to see that $A \wedge B \implies C$:

$$\hat{\mu}^m_{i^*} > \mu_{i^*} - \Delta/2 \geq \mu_j + \Delta - \Delta/2 = \mu_j + \Delta/2 > \hat{\mu}^m_j.$$

This implies that:

$$\mathbb{P}(i^*_m = j) \leq \mathbb{P}(\hat{\mu}^m_j \geq \hat{\mu}^m_{i^*}) = \mathbb{P}(\overline{C}) \leq \mathbb{P}(\overline{A} \vee \overline{B}) \leq \mathbb{P}(\overline{A}) + \mathbb{P}(\overline{B}).$$

It remains to show that $\mathbb{P}(\overline{A}) \leq \delta_m/4$ and $\mathbb{P}(\overline{B}) \leq \delta_m/4$:

$$\begin{aligned} \mathbb{P}(\overline{A}) &= \mathbb{P}(\hat{\mu}^m_{i^*} \leq \mu_{i^*} - \Delta/2) \\ &\leq \exp\left(-\frac{1}{2}(\Delta/2)^2 \left\lceil \frac{\log(4/\delta_m)}{2^{-3}\Delta^2} \right\rceil\right) \\ &\leq \exp\left(-\frac{1}{2}(\Delta/2)^2 \frac{\log(4/\delta_m)}{2^{-3}\Delta^2}\right) \\ &= \delta_m/4. \end{aligned}$$

Similarly, $\mathbb{P}(\overline{B}) \leq \delta_m/4$. □

**Lemma 9.** *The sample complexity of Algorithm 4 is*

$$\mathcal{O}\left(\log n + \frac{1}{\Delta^2}\log\frac{n}{\delta}\right).$$

*Proof.* The number of arms sampled in the $\ell$-th round is $3 \cdot n_\ell$, so the number of arm sampled in total:

$$\begin{aligned}
\sum_{\ell=1}^{m} 3 \cdot n_\ell &\leq 3\sum_{\ell=1}^{m} \left\lceil \log\left(4/\delta_\ell\right) / \left(2^{2m-2\ell-5}\Delta^2\right)\right\rceil \\
&\leq 3\,m + 3\sum_{\ell=1}^{m} \log\left(4/\delta_\ell\right) / \left(2^{2m-2\ell-5}\Delta^2\right) \\
&= 3\,m + 3\sum_{\ell=1}^{m} \log\left(2^{\ell+2}/\delta\right) / \left(2^{2m-2\ell-5}\Delta^2\right) \\
&= 3\,m + \frac{3}{2^{2m-5}\Delta^2}\left(\sum_{\ell=1}^{m} 2^{2\ell}\log\left(2^{\ell+2}/\delta\right)\right) \\
&\leq 3\,m + \frac{3}{2^{2m-5}\Delta^2}\log\left(2^{m+2}/\delta\right)\sum_{\ell=1}^{m} 2^{2\ell} \\
&\leq 3\,m + \frac{3}{2^{2m-5}\Delta^2}\log\left(2^{m+2}/\delta\right)\frac{2^{2m+2}}{3} \\
&= 3\,m + 128\Delta^{-2}\log\left(2^{m+2}/\delta\right) \\
&= \mathcal{O}\left(\log n + \frac{1}{\Delta^2}\log\frac{n}{\delta}\right).
\end{aligned}$$

□

*Proof of Theorem 8.* By Lemma 7 the probability that the optimal arm is removed in the first $m-1$ rounds is less than or equal to

$$\sum_{\ell=1}^{m-1}\delta_\ell = \sum_{\ell=1}^{m-1}\frac{\delta}{2^\ell} = \delta - \frac{\delta}{2^{m-1}}.$$

By Lemma 8, the probability that the algorithm doesn't return the best of the three arms in the last round is less than or equal to $\delta_m = \delta/2^m$, so the probability, that the arm returned by the algorithm is not optimal is less than $\delta$. By Lemma 9, the sample complexity of the algorithm is

$$\mathcal{O}\left(\log n + \frac{1}{\Delta^2}\log\frac{n}{\delta}\right).$$

□

**Remark 2.** *In the general case when* $2^{m-1}+1 < n \leq 2^m+1$ *we can do the following: At first update the indices:*

$$i \leftarrow i + \left\lfloor \frac{2^m + 1 - n}{2} \right\rfloor .$$

*This way we can sample the arms* $a_j$, $j \in \{i \cdot 2^{m-2}, i = 1, 2, 3\}$. *Sample all of them* $n_1 = \lceil \log(8/\delta)/(2^{2m-7}\Delta^2) \rceil$ *times. Let* $\hat{\mu}_j^1$ *denote their empirical values and let* $i_1^* = \arg\max_j \hat{\mu}_j^1$. *Keep the* $2^{m-1}+1$ *arms closest to the arm* $a_{i_1^*}$, *denote the set of these arms with* $S_2$. *Renumber the arms from* 0 *to* $2^{m-1}$. *Set* $\delta_2 = \delta/4$ *and* $\ell = 2$. *After this we can continue with the second round of Algorithm 4.*

## 5.2 Experiments

Based on Lemma 9, the sample complexity of the algorithm is

$$\mathcal{O}\left(\log n + \frac{1}{\Delta^2} \log \frac{n}{\delta}\right).$$

The following plots illustrate the relationship between the sample complexity and certain parameters of Algorithm 4. Figure 1 demonstrates the logarithmic relationship between the number of arms and the sample complexity. Figure 2 explores the dependence of sample complexity on $\delta$. We can see, that as $\delta$ increases, the sample complexity decreases, which aligns with the derived sample complexity formula. On Figure 3 we can observe the inverse relationship between the sample complexity and $\Delta$ as higher values of $\Delta$ correspond to lower sample complexities, which is consistent with the $\frac{1}{\Delta^2}$ term in the derived formula of the sample complexity.

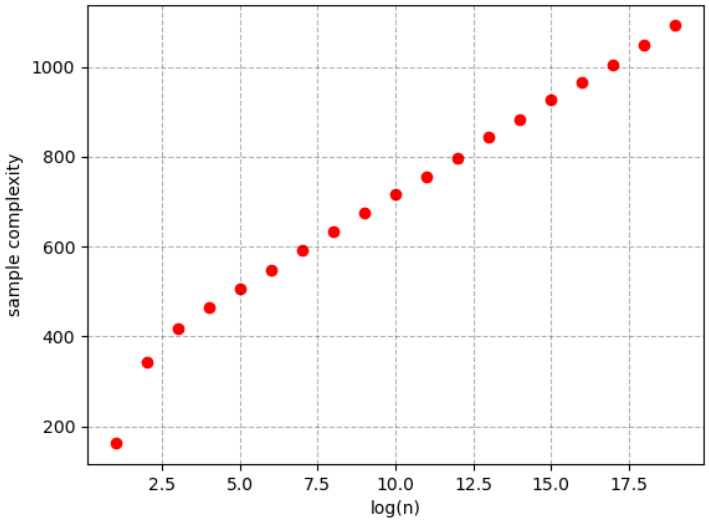



Figure 1: The relationship between the number of arms ($n$) and the sample complexity in case of $\Delta = 1$, $\delta = 0.01$.

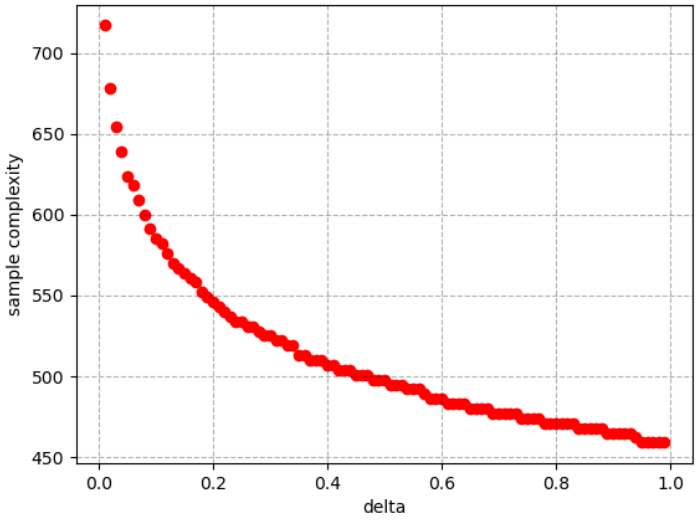



Figure 2: The relationship between the $\delta$ parameter and the sample complexity in case of $\Delta = 1$, $n = 1025$.

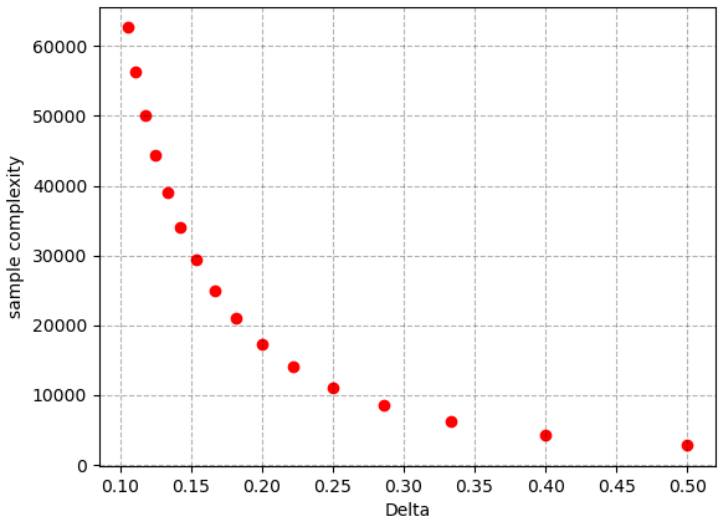



Figure 3: The relationship between the $\Delta$ parameter and the sample complexity in case of $\delta = 0.01$, $n = 1025$.

# 6 Infinitely Many Arms with a Concave Structure

In this section we will consider the problem in which the arms are the points of the $[0, 1]$ interval and an unknown concave function describes the expectations of the arms. The assumptions on the arms are the following:

**Assumption 5.** *The arms are the points of the* $[0, 1]$ *interval and the expectation of arm* $x \in [0, 1]$ *is* $f(x)$*, where* $f : [0, 1] \to \mathbb{R}$ *is an unknown concave function.*

**Assumption 6.** *The arms are* 1*-subgaussian.*

Under these assumptions we can create an $(\varepsilon, \delta)$-PAC algorithm, which is more efficient than Median Elimination.

## 6.1 Algorithm

In each round of this new algorithm, either the set of the arms is reduced or the algorithm terminates. In round $\ell$ the set of arms is denoted by $S_\ell$, which is divided into four equal-length subintervals by five arms: $x_0^\ell$, $x_1^\ell$, $x_2^\ell$, $x_3^\ell$, $x_4^\ell$ (Figure 4). The expectation of arm $x_i^\ell$ is denoted by $\mu_i^\ell$. In round $\ell$ we already have estimations of $\mu_0^\ell$ and $\mu_4^\ell$ from the previous round. We want to estimate $\mu_1^\ell, \mu_2^\ell$ and $\mu_3^\ell$ as well, so we sample $x_1^\ell$, $x_2^\ell$, $x_3^\ell$ many times and we estimate the expectations with the sample means. The sample mean of arm $x_i^\ell$ is denoted by $\hat{\mu}_i^\ell$.

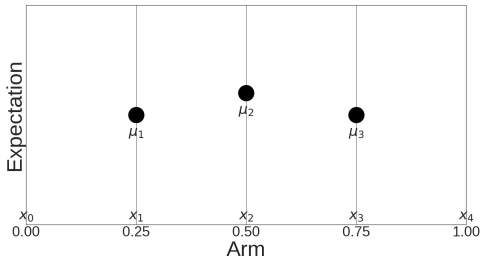



Figure 4: The set of arms is divided into four equal-length subintervals

If $\hat{\mu}_1^\ell$ and $\hat{\mu}_3^\ell$ are close to $\hat{\mu}_2^\ell$, then $\mu_1^\ell$ and $\mu_3^\ell$ are close to $\mu_2^\ell$. If

$$\mu_2^\ell - \varepsilon/2 \leq \mu_1^\ell, \mu_3^\ell \leq \mu_2^\ell + \varepsilon/2, \qquad (*)$$

then the arm $x_2^\ell$ will be $\varepsilon$-optimal because of the concavity and the algorithm can return arm $x_2^\ell$:

- An arm with expectation larger than $\mu_2^\ell + \varepsilon$ cannot be between $x_0^\ell$ and $x_1^\ell$, because in this case $\mu_1^\ell$ would be under the line segment connecting the expectation of this arm with $\mu_2^\ell$ and this cannot happen because of the concavity (Figure 5).
- An arm with expectation larger than $\mu_2^\ell + \varepsilon$ cannot be between $x_1^\ell$ and $x_2^\ell$, because in this case $\mu_2^\ell$ would be under the line segment connecting the expectation of this arm with $\mu_3^\ell$ and again, this cannot happen because of the concavity (Figure 6).

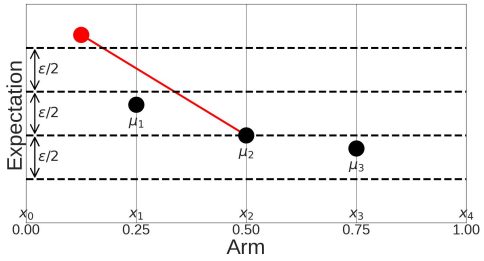



Figure 5: An arm with expectation larger than $\mu_2^\ell + \varepsilon$ cannot be between $x_0^\ell$ and $x_1^\ell$ if (*) holds

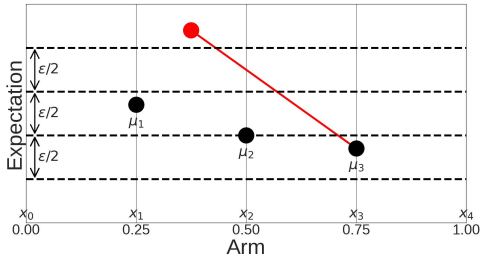



Figure 6: If (*) holds, an arm with expectation larger than $\mu_2^\ell + \varepsilon$ cannot be between $x_1^\ell$ and $x_2^\ell$

- Similarly, neither the $(x_2^\ell, x_3^\ell)$ nor the $(x_3^\ell, x_4^\ell)$ interval can contain an arm with expectation larger than $\mu_2^\ell + \varepsilon$.

If $\hat{\mu}_1^\ell$ is much smaller than $\hat{\mu}_2^\ell$, then $\mu_1^\ell \leq \mu_2^\ell$ and the optimal arm cannot be on the left of arm $x_1^\ell$ otherwise $\mu_1^\ell$ would be under the line segment connecting the expectation of the optimal arm with $\mu_2^\ell$, which contradicts the fact that a concave function describes the expectations (Figure 7). So in this case we can remove the arms on the left of arm $x_1^\ell$.

Similarly, if $\hat{\mu}_1^\ell$ is much larger than $\hat{\mu}_2^\ell$, then $\mu_1^\ell \geq \mu_2^\ell$ and because of the concavity, the optimal arm cannot be on the right of arm $x_2^\ell$. This way we can remove the arms on the right of arm $x_2^\ell$.

We can also remove arms similarly, if $\hat{\mu}_3^\ell$ is much smaller or larger than $\hat{\mu}_2^\ell$.

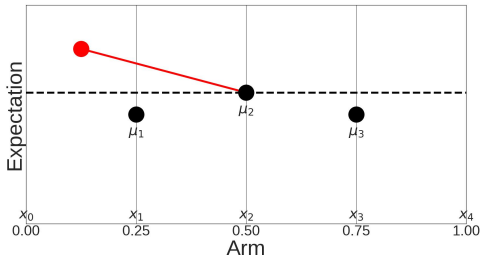



Figure 7: If $\mu_1^\ell \leq \mu_2^\ell$, the optimal arm cannot be on the left of arm $x_1^\ell$

**Algorithm 5**

**Input:** $\delta > 0$, $\varepsilon > 0$

**Output:** an arm which is $\varepsilon$-optimal with probability at least $1 - \delta$

Set $\delta_0 = \delta/2$, $x_0^1 = 0$, $x_1^1 = 0.25$, $x_2^1 = 0.5$, $x_3^1 = 0.75$, $x_4^1 = 1$.

Sample $n_0 = \lceil 128 \log(4/\delta_0)/\varepsilon^2 \rceil$ times $x_0^1$ and $x_4^1$.

Let $\hat{\mu}_0^1$, $\hat{\mu}_4^1$ denote the sample means and $\mu_0^1$, $\mu_4^1$ denote the expectations.

Set $S_1 = [0, 1]$, $\delta_1 = \delta_0/2$, $\ell = 1$.

**while TRUE do**

$S_{\ell+1} = S_\ell$.

Sample $n_\ell = \lceil 128 \log(6/\delta_\ell)/\varepsilon^2 \rceil$ times the three arms: $x_1^\ell$, $x_2^\ell$, $x_3^\ell$.

Let $\hat{\mu}_1^\ell, \hat{\mu}_2^\ell, \hat{\mu}_3^\ell$ and $\mu_1^\ell, \mu_2^\ell, \mu_3^\ell$ denote the sample means and expectations.

**if** $\hat{\mu}_1^\ell, \hat{\mu}_3^\ell \in (\hat{\mu}_2^\ell - \varepsilon/4, \hat{\mu}_2^\ell + \varepsilon/4)$ **then return** $x_2^\ell$

**if** $\hat{\mu}_1^\ell \geq \hat{\mu}_2^\ell + \varepsilon/4$ **then** $S_{\ell+1} = S_{\ell+1} \setminus (x_2^\ell, x_4^\ell]$

**if** $\hat{\mu}_1^\ell \leq \hat{\mu}_2^\ell - \varepsilon/4$ **then** $S_{\ell+1} = S_{\ell+1} \setminus [x_0^\ell, x_1^\ell)$

**if** $\hat{\mu}_3^\ell \geq \hat{\mu}_2^\ell + \varepsilon/4$ **then** $S_{\ell+1} = S_{\ell+1} \setminus [x_0^\ell, x_2^\ell)$

**if** $\hat{\mu}_3^\ell \leq \hat{\mu}_2^\ell - \varepsilon/4$ **then** $S_{\ell+1} = S_{\ell+1} \setminus (x_3^\ell, x_4^\ell]$

$x_0^{\ell+1} = \min S_{\ell+1}$, $x_4^{\ell+1} = \max S_{\ell+1}$,

$x_1^{\ell+1} = \frac{3}{4} \cdot x_0^{\ell+1} + \frac{1}{4} \cdot x_4^{\ell+1}$, $x_2^{\ell+1} = \frac{1}{2} \cdot x_0^{\ell+1} + \frac{1}{2} \cdot x_4^{\ell+1}$, $x_3^{\ell+1} = \frac{1}{4} \cdot x_0^{\ell+1} + \frac{3}{4} \cdot x_4^{\ell+1}$.

Let $\hat{\mu}_0^{\ell+1}, \hat{\mu}_4^{\ell+1}$ denote the sample means of $x_0^{\ell+1}$ and $x_4^{\ell+1}$ calculated in round $\ell$, and let $\mu_0^{\ell+1}$ and $\mu_4^{\ell+1}$ denote the expectations of $x_0^{\ell+1}$ and $x_4^{\ell+1}$.

$\delta_{\ell+1} = \delta_\ell/2$, $\ell = \ell + 1$

**end while**

**Theorem 9.** *Under Assumptions 5-6, Algorithm 5 finds an $\varepsilon$-optimal arm with probability at least $1 - \delta$.*

First, we will prove a few lemmas, which will be used for the proof of the theorem.

**Lemma 10.**

$$\mathbb{P}(\exists \ell,\, \exists i \in \{1, 2, 3\} : |\hat{\mu}_i^\ell - \mu_i^\ell| \geq \varepsilon/8) \leq \delta/2,$$

$$\mathbb{P}(|\hat{\mu}_0^1 - \mu_0^1| \geq \varepsilon/8 \vee |\hat{\mu}_4^1 - \mu_4^1| \geq \varepsilon/8) \leq \delta/2.$$

*Proof.* For every phase $\ell \geq 1$:

$$\begin{aligned}
&\mathbb{P}(\exists i \in \{1,2,3\} : |\hat{\mu}_i^\ell - \mu_i^\ell| \geq \varepsilon/8) \\
&\quad \leq \sum_{i=1}^{3} \mathbb{P}(|\hat{\mu}_i^\ell - \mu_i^\ell| \geq \varepsilon/8) \\
&\quad \leq 3 \cdot 2 \exp\left(-n_\ell \cdot (\varepsilon/8)^2/2\right) \\
&\quad \leq \delta_\ell.
\end{aligned}$$

$$\begin{aligned}
&\mathbb{P}(\exists \ell,\, \exists i \in \{1,2,3\} : |\hat{\mu}_i^\ell - \mu_i^\ell| \geq \varepsilon/8) \\
&\quad \leq \sum_{\ell=1}^{\infty} \mathbb{P}(\exists i \in \{1,2,3\} : |\hat{\mu}_i^\ell - \mu_i^\ell| \geq \varepsilon/8) \\
&\quad \leq \sum_{\ell=1}^{\infty} \delta_\ell = \delta/2.
\end{aligned}$$

$$\begin{aligned}
&\mathbb{P}(|\hat{\mu}_0^1 - \mu_0^1| \geq \varepsilon/8 \vee |\hat{\mu}_0^4 - \mu_0^4| \geq \varepsilon/8) \\
&\quad \leq \mathbb{P}(|\hat{\mu}_0^1 - \mu_0^1| \geq \varepsilon/8) + \mathbb{P}(|\hat{\mu}_0^4 - \mu_0^4| \geq \varepsilon/8) \\
&\quad \leq 2 \cdot 2 \exp\left(-n_0 \cdot (\varepsilon/8)^2/2\right) \\
&\quad \leq \delta_0 = \delta/2.
\end{aligned}$$

□

**Lemma 11.** *Suppose that $|\hat{\mu}_i^\ell - \mu_i^\ell| \leq \varepsilon/8$ for $i = 0,1,2,3,4$. Let $x^*$ denote the optimal arm. If $i < j$ and $\hat{\mu}_i^\ell \geq \hat{\mu}_j^\ell + \varepsilon/4$, then $x^* \leq x_j^\ell$. Similarly, if $i > j$ and $\hat{\mu}_i^\ell \geq \hat{\mu}_j^\ell + \varepsilon/4$, then $x^* \geq x_j^\ell$.*

*Proof.* If $\hat{\mu}_i^\ell \geq \hat{\mu}_j^\ell + \varepsilon/4$ then $\mu_i^\ell \geq \mu_j^\ell$:

$$\mu_i^\ell \geq \hat{\mu}_i^\ell - \varepsilon/8 \geq \hat{\mu}_j^\ell + \varepsilon/4 - \varepsilon/8 = \hat{\mu}_j^\ell + \varepsilon/8 \geq \mu_j^\ell.$$

Arguing indirectly, assume that $x^* > x_j^\ell$. Then there exists a $t \in [0,1]$ such that $x_j^\ell = t \cdot x_i^\ell + (1-t) \cdot x^*$. Because of the concavity:

$$f(x_j^\ell) = f(t \cdot x_i^\ell + (1-t) \cdot x^*) \geq t \cdot f(x_i^\ell) + (1-t) \cdot f(x^*) > f(x_i^\ell).$$

It contradicts the fact that $\mu_i^\ell \geq \mu_j^\ell$. The other case can be proven similarly. □

**Lemma 12.** *If* $|\hat{\mu}_i^\ell - \mu_i^\ell| \leq \varepsilon/8$ *for* $i = 0, 1, 2, 3, 4$ *and* $\hat{\mu}_{i-1}^\ell, \hat{\mu}_{i+1}^\ell \in (\hat{\mu}_i^\ell - \varepsilon/4, \hat{\mu}_i^\ell + \varepsilon/4)$, *then*

$$\mu_{i-1}^\ell, \mu_{i+1}^\ell \in [\mu_i^\ell - \varepsilon/2, \mu_i^\ell + \varepsilon/2].$$

*Proof.* Using the assumptions of the lemma, we get the following:

$$\mu_{i-1}^\ell \geq \hat{\mu}_{i-1}^\ell - \varepsilon/8 \geq \hat{\mu}_i^\ell - \varepsilon/4 - \varepsilon/8 = \hat{\mu}_i^\ell - 3/8 \cdot \varepsilon \geq \mu_i^\ell - \varepsilon/8 - 3/8 \cdot \varepsilon = \mu_i^\ell - \varepsilon/2,$$

$$\mu_{i-1}^\ell \leq \hat{\mu}_{i-1}^\ell + \varepsilon/8 \leq \hat{\mu}_i^\ell + \varepsilon/4 + \varepsilon/8 = \hat{\mu}_i^\ell + 3/8 \cdot \varepsilon \leq \mu_i^\ell + \varepsilon/8 + 3/8 \cdot \varepsilon = \mu_i^\ell + \varepsilon/2.$$

Similarly, $\mu_{i+1}^\ell \in [\mu_i^\ell - \varepsilon/2, \mu_i^\ell + \varepsilon/2]$.

□

**Lemma 13.** *If* $|\hat{\mu}_i^\ell - \mu_i^\ell| \leq \varepsilon/8$ *for* $i = 1, 2, 3$ *and* $\hat{\mu}_1^\ell, \hat{\mu}_3^\ell \in (\hat{\mu}_2^\ell - \varepsilon/4, \hat{\mu}_2^\ell + \varepsilon/4)$, *then*

$$\mu_2^\ell \geq \max_{x \in S_\ell} f(x) - \varepsilon.$$

*Proof.* Let $x^* = \arg\max_{x \in S_\ell} f(x)$. Arguing indirectly assume, that $f(x^*) > f(x_2^\ell) + \varepsilon$. If $x^* < x_1^\ell$, then there exists a $t \in [0, 1]$ such that $x_1^\ell = (1 - t) \cdot x^* + t \cdot x_2^\ell$. The fact that $x_1^\ell$ is closer to $x^*$ than to $x_2^\ell$ implies that $t \leq 1/2$. Because of the concavity:

$$\begin{aligned} f(x_1^\ell) &= f((1 - t) \cdot x^* + t \cdot x_2^\ell) \\ &\geq (1 - t) \cdot f(x^*) + t \cdot f(x_2^\ell) \\ &> (1 - t) \cdot (f(x_2^\ell) + \varepsilon) + t \cdot f(x_2^\ell) \\ &= f(x_2^\ell) + (1 - t) \cdot \varepsilon \\ &\geq f(x_2^\ell) + \varepsilon/2. \end{aligned}$$

This contradicts the fact that $f(x_1^\ell) \leq f(x_2^\ell) + \varepsilon/2$.
If $x_1^\ell < x^* < x_2^\ell$, then there exists a $t \in [0, 1]$ such that $x_2^\ell = (1 - t) \cdot x^* + t \cdot x_3^\ell$. The point $x_2^\ell$ is closer to $x^*$ than to $x_3^\ell$, so $t \leq 1/2$. Because of the concavity:

$$\begin{aligned} f(x_2^\ell) &= f((1 - t) \cdot x^* + t \cdot x_3^\ell) \\ &\geq (1 - t) \cdot f(x^*) + t \cdot f(x_3^\ell) \\ &> (1 - t) \cdot (f(x_2^\ell) + \varepsilon) + t \cdot (f(x_2^\ell) - \varepsilon/2) \\ &= f(x_2^\ell) + (1 - 3/2 \cdot t) \cdot \varepsilon \\ &> f(x_2^\ell). \end{aligned}$$

This is again a contradiction.
If $x_2^\ell < x^* < x_3^\ell$, then there exists a $t \in [0,1]$ such that $x_2^\ell = (1-t) \cdot x_1^\ell + t \cdot x^*$. The fact that $x_2^\ell$ is closer to $x^*$ than to $x_1^\ell$ implies that $1 > t \geq 1/2$. Because of the concavity:

$$\begin{aligned} f(x_2^\ell) &= f((1-t) \cdot x_1^\ell + t \cdot x^*) \\ &\geq (1-t) \cdot f(x_1^\ell) + t \cdot f(x^*) \\ &> (1-t) \cdot (f(x_2^\ell) - \varepsilon/2) + t \cdot (f(x_2^\ell) + \varepsilon) \\ &= f(x_2^\ell) + (3/2 \cdot t - 1/2) \cdot \varepsilon \\ &> f(x_2^\ell). \end{aligned}$$

Which is again a contradiction.
If $x_3^\ell < x^*$, then there exists a $t \in [0,1]$ such that $x_3^\ell = (1-t) \cdot x_2^\ell + t \cdot x^*$. The point $x_3^\ell$ is closer to $x^*$ than to $x_2^\ell$, so $t \geq 1/2$. Because of the concavity:

$$\begin{aligned} f(x_3^\ell) &= f((1-t) \cdot x_2^\ell + t \cdot x^*) \\ &\geq (1-t) \cdot f(x_2^\ell) + t \cdot f(x^*) \\ &> (1-t) \cdot f(x_2^\ell) + t \cdot (f(x_2^\ell) + \varepsilon) \\ &= f(x_2^\ell) + t \cdot \varepsilon \\ &\geq f(x_2^\ell) + \varepsilon/2. \end{aligned}$$

This contradicts the fact that $f(x_3^\ell) \leq f(x_2^\ell) + \varepsilon/2$.

□

*Proof of Theorem.* By Lemma 10, $\mathbb{P}(\exists \ell, \exists i \in \{0,1,2,3,4\} : |\hat{\mu}_i^\ell - \mu_i^\ell| \geq \varepsilon/8) \leq \delta$. Now consider the case, when $\forall \ell, \forall i \in \{0,1,2,3,4\} : |\hat{\mu}_i^\ell - \mu_i^\ell| \leq \varepsilon/8$. In each round at least a quarter of the arms are removed or the algorithm terminates. By Lemma 16 the optimal arm is never discarded and by Lemma 13 an $\varepsilon$-optimal arm is returned when the algorithm terminates. This way Algorithm 5 returns an $\varepsilon$-optimal arm with probability at least $1 - \delta$. □

## 6.2 Experiments

The following plots demonstrate the relationship between the sample complexity of Algorithm 5 and the input parameters of the algorithm. The $f(x) = -x^2$ function describes the expectations of the arms in both cases. Figure 8 demonstrates the inverse relationship between the sample complexity and $\varepsilon$. As $\varepsilon$ increases, the sample complexity steeply decreases. Figure 9 explores the dependence of sample complexity on $\delta$. We can see, that the sample complexity decreases much more slowly with an increase in the $\delta$ parameter compared to an increase in the $\varepsilon$ parameter.

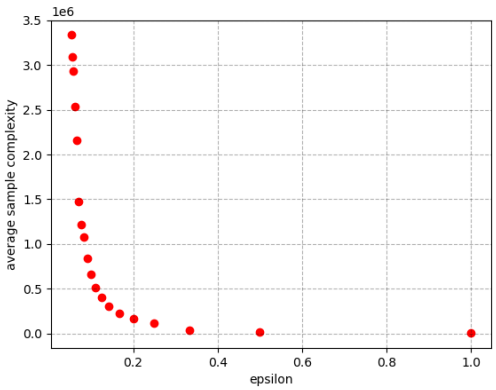



Figure 8: The relationship between $\varepsilon$ and the sample complexity in case of $\delta = 0.05$

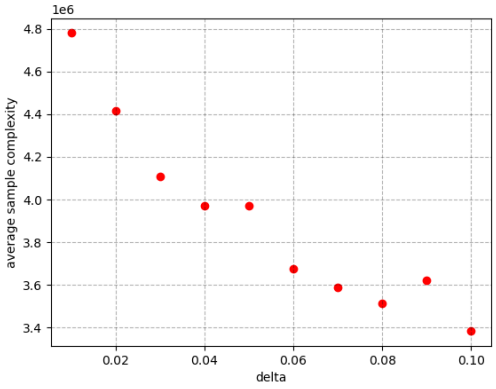



Figure 9: The relationship between $\delta$ and the sample complexity in case of $\varepsilon = 0.05$

# 7 Lipschitz Continuous Case

In this section we will consider the case when there are infinitely many arms, which are the points of the $(0, 1)$ interval, and an unknown concave function describes the expectations of the arms, where this function is Lipschitz continuous with a known Lipschitz constant:

**Assumption 7.** *Function $f$ is Lipschitz continuous with Lipschitz constant $L$:*

$$|f(x) - f(y)| \leq L \cdot |x - y| \quad \forall\, x, y \in [0, 1].$$

In this case, if the length of the set of arms is less than or equal to $2 \cdot \varepsilon/L$ and it contains the optimal arm, then the arm in the middle of the interval will be $\varepsilon$-optimal. By modifying Algorithm 5 so that it returns the arm in the middle when the length of the set of arms is less than or equal to $2 \cdot \varepsilon/L$, we get Algorithm 6.

## 7.1 Algorithm

**Theorem 10.** *Under Assumptions 5-7, Algorithm 6 is an $(\varepsilon, \delta)$-PAC algorithm, and its sample complexity is*

$$\mathcal{O}\left(\frac{1}{\varepsilon^2}\left(\left(\log\frac{L}{\varepsilon}\right)^2 + \left(\log\frac{L}{\varepsilon}\right) \cdot \log\frac{1}{\delta}\right)\right).$$

*Proof.* If $f$ is Lipschitz continuous with Lipschitz constant $L$ and there is an interval with length less than or equal to $2 \cdot \varepsilon/L$ containing the optimal arm $(y)$, then the arm in the middle of the interval $(x)$ is $\varepsilon$-optimal:

$$|x - y| \leq \varepsilon/L \implies |f(x) - f(y)| \leq \varepsilon.$$

By Theorem 9, the probability that the optimal arm is removed by Algorithm 5 is less than $\delta$. This way Algorithm 6 is an $(\varepsilon, \delta)$-PAC algorithm.
After $\ell$ rounds the length of the interval is $(1/2)^\ell$, so $\ell = \left\lfloor \log_2 \frac{L}{\varepsilon} \right\rfloor$ round is enough to get an interval no longer than $2 \cdot \varepsilon/L$. So the algorithm terminates in $\ell$ rounds and the number of samples required by the algorithm is bounded by:

$$
\begin{aligned}
&2\left\lceil\frac{128}{\varepsilon^2}\log\frac{4}{\delta_0}\right\rceil+3\sum_{i=1}^{\ell}\left\lceil\frac{128}{\varepsilon^2}\log\frac{6}{\delta_i}\right\rceil\\
&\quad\le 3\ell+2+2\cdot\frac{128}{\varepsilon^2}\log\frac{4}{\delta_0}+3\cdot\frac{128}{\varepsilon^2}\sum_{i=1}^{\ell}\log\frac{6}{\delta_i}\\
&\quad\le 3\ell+2+3\cdot\frac{128}{\varepsilon^2}\sum_{i=0}^{\ell}\log\frac{6}{\delta_i}\\
&\quad\le 3\ell+2+3\cdot\frac{128}{\varepsilon^2}\sum_{i=1}^{\ell}\log\frac{12\cdot 2^i}{\delta}\\
&\quad= 3\ell+2+3\cdot\frac{128}{\varepsilon^2}\log\left(\prod_{i=1}^{\ell}\frac{12\cdot 2^i}{\delta}\right)\\
&\quad= 3\ell+2+3\cdot\frac{128}{\varepsilon^2}\log\left(\frac{12^{\ell+1}}{\delta^{\ell+1}}\cdot 2^{\frac{\ell(\ell+1)}{2}}\right)\\
&\quad= 3\ell+2+3\cdot\frac{128}{\varepsilon^2}\left((\ell+1)\log\frac{12}{\delta}+\frac{\ell(\ell+1)}{2}\log 2\right)\\
&\quad= \mathcal{O}\left(\frac{1}{\varepsilon^2}\left(\ell^2+\ell\cdot\log\frac{1}{\delta}\right)\right)\\
&\quad= \mathcal{O}\left(\frac{1}{\varepsilon^2}\left(\left(\log\frac{L}{\varepsilon}\right)^2+\left(\log\frac{L}{\varepsilon}\right)\cdot\log\frac{1}{\delta}\right)\right).
\end{aligned}
$$

□

**Algorithm 6**

**Input:** $\varepsilon > 0$, $\delta > 0$

**Output:** an arm which is $\varepsilon$-optimal with probability at least $1 - \delta$

Set $\delta_0 = \delta/2$, $x_0^1 = 0$, $x_1^1 = 0.25$, $x_2^1 = 0.5$, $x_3^1 = 0.75$, $x_4^1 = 1$.

Sample $n_0 = \lceil 128 \log(4/\delta_0)/\varepsilon^2 \rceil$ times $x_0^1$ and $x_4^1$.

Let $\hat{\mu}_0^1$, $\hat{\mu}_4^1$ denote the sample means and $\mu_0^1$, $\mu_4^1$ denote the expectations.

Set $S_1 = [0, 1]$, $\delta_1 = \delta_0/2$, $\ell = 1$.

**while TRUE do**

**if** $|S_\ell| \leq 2 \cdot \varepsilon/L$ **then return** $x_2^\ell$

$S_{\ell+1} = S_\ell$.

Sample $n_\ell = \lceil 128 \log(6/\delta_\ell)/\varepsilon^2 \rceil$ times the three arms: $x_1^\ell$, $x_2^\ell$, $x_3^\ell$.

Let $\hat{\mu}_1^\ell, \hat{\mu}_2^\ell, \hat{\mu}_3^\ell$ and $\mu_1^\ell, \mu_2^\ell, \mu_3^\ell$ denote the sample means and expectations.

**if** $\hat{\mu}_1^\ell, \hat{\mu}_3^\ell \in (\hat{\mu}_2^\ell - \varepsilon/4, \hat{\mu}_2^\ell + \varepsilon/4)$ **then return** $x_2^\ell$

**if** $\hat{\mu}_1^\ell \geq \hat{\mu}_2^\ell + \varepsilon/4$ **then** $S_{\ell+1} = S_{\ell+1} \setminus (x_2^\ell, x_4^\ell]$

**if** $\hat{\mu}_1^\ell \leq \hat{\mu}_2^\ell - \varepsilon/4$ **then** $S_{\ell+1} = S_{\ell+1} \setminus [x_0^\ell, x_1^\ell)$

**if** $\hat{\mu}_3^\ell \geq \hat{\mu}_2^\ell + \varepsilon/4$ **then** $S_{\ell+1} = S_{\ell+1} \setminus [x_0^\ell, x_2^\ell)$

**if** $\hat{\mu}_3^\ell \leq \hat{\mu}_2^\ell - \varepsilon/4$ **then** $S_{\ell+1} = S_{\ell+1} \setminus (x_3^\ell, x_4^\ell]$

$x_0^{\ell+1} = \min S_{\ell+1}$, $x_4^{\ell+1} = \max S_{\ell+1}$,

$x_1^{\ell+1} = \frac{3}{4} \cdot x_0^{\ell+1} + \frac{1}{4} \cdot x_4^{\ell+1}$, $x_2^{\ell+1} = \frac{1}{2} \cdot x_0^{\ell+1} + \frac{1}{2} \cdot x_4^{\ell+1}$, $x_3^{\ell+1} = \frac{1}{4} \cdot x_0^{\ell+1} + \frac{3}{4} \cdot x_4^{\ell+1}$.

Let $\hat{\mu}_0^{\ell+1}, \hat{\mu}_4^{\ell+1}$ denote the sample means of $x_0^{\ell+1}$ and $x_4^{\ell+1}$ calculated in round $\ell$, and let $\mu_0^{\ell+1}$ and $\mu_4^{\ell+1}$ denote the expectations of $x_0^{\ell+1}$ and $x_4^{\ell+1}$.

$\delta_{\ell+1} = \delta_\ell/2$, $\ell = \ell + 1$

**end while**

## 7.2 Experiments

Based on Theorem 10, the sample complexity of the algorithm is

$$\mathcal{O}\left(\frac{1}{\varepsilon^2}\left(\left(\log\frac{L}{\varepsilon}\right)^2 + \left(\log\frac{L}{\varepsilon}\right) \cdot \log\frac{1}{\delta}\right)\right).$$

The following plots demonstrate the relationship between the sample complexity and the $\varepsilon$ and $\delta$ parameters of the algorithm when $f(x) = |x - 0.5|$ describes

the expectations of the arms and $L = 1$. Figure 10 explores the dependence of sample complexity on $\varepsilon$, the desired precision of the estimate. We can see, that as $\varepsilon$ increases, the sample complexity decreases steeply, this aligns with the derived sample complexity formula. Figure 11 reveals the inverse relationship between the sample complexity and $\delta$. We can observe that higher values of $\delta$ correspond to lower sample complexities. This is consistent with the $\log \frac{1}{\delta}$ term in the formula of the sample complexity.

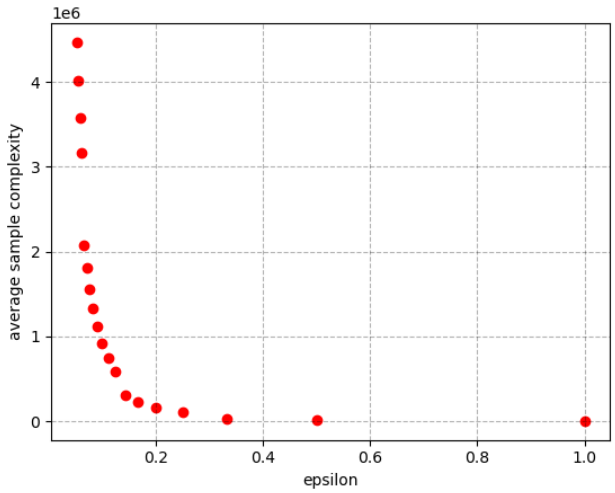



Figure 10: The relationship between $\varepsilon$ and the sample complexity in case of $\delta = 0.05$

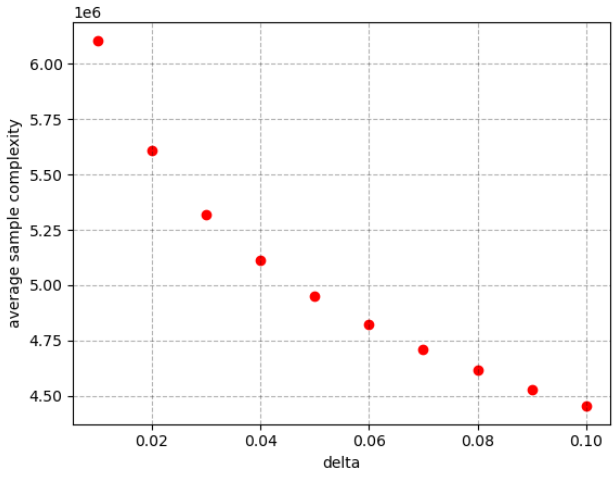



Figure 11: The relationship between $\delta$ and the sample complexity in case of $\varepsilon = 0.05$

Figure 12 shows the relationship between the sample complexity and the Lipschitz constant used by the algorithm when the $f(X) = -100\,|x - 0.5|$ function describes the expectations of the arms. We can observe the logarithmic relationship between $L$ and the sample complexity suggested by the Theorem 10.

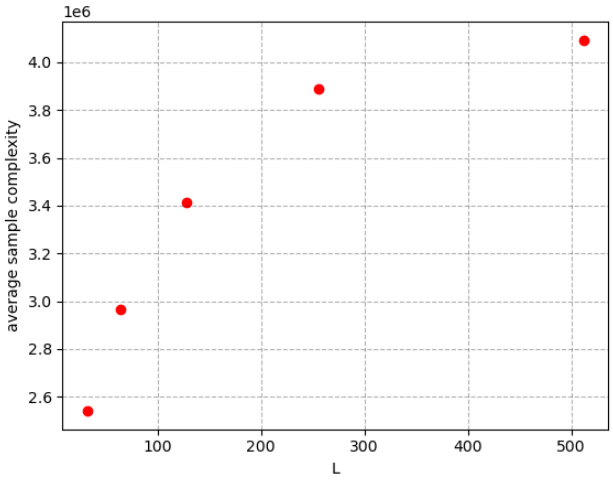



Figure 12: The relationship between the sample complexity and the Lipschitz constant ($L$) used by the algorithm in case of $\varepsilon = 0.1$ and $\delta = 0.1$

# 8 Finitely Many Arms with a Concave Structure

In this section we will consider the case when there are finitely many arms and an unknown concave function describes the expectations of the arms. By modifying Algorithm 5 we can achieve to halve the set of arms in each round. First we will see that this modified algorithm can be used to solve the case when there are $2^m + 1$ arms for some $m \in \mathbb{N}$ and then the general case can be solved using this special case. The assumptions on the arms are the following:

**Assumption 8.** *There are $n = 2^m + 1$, $m \geq 1$ arms numbered from 0 to $2^m$.*

**Assumption 9.** *The expectation of arm $i$ is $f(i)$, where $f : \mathbb{R} \to \mathbb{R}$ is an unknown concave function.*

**Assumption 10.** *The arms are 1-subgaussian.*

The difference between Algorithm 5 and Algorithm 7 is that in the case when $\hat{\mu}_1^\ell \in (\hat{\mu}_2^\ell - \varepsilon/4, \hat{\mu}_2^\ell + \varepsilon/4)$ and $\hat{\mu}_3^\ell \leq \hat{\mu}_2^\ell - \varepsilon/4$ then in Algorithm 5 only a quarter of the arms is removed, while in Algorithm 7 the arm $x_{1.5}^\ell = (x_1^\ell + x_2^\ell)/2$ is sampled and based on the results another quarter of the arms is removed. Similarly, when $\hat{\mu}_1^\ell \leq \hat{\mu}_2^\ell - \varepsilon/4$ and $\hat{\mu}_3^\ell \in (\hat{\mu}_2^\ell - \varepsilon/4, \hat{\mu}_2^\ell + \varepsilon/4)$ then arm $x_{2.5}^\ell = (x_2^\ell + x_3^\ell)/2$ is sampled in order to remove another quarter of the arms.

## 8.1 Algorithm

**Theorem 11.** *Under Assumptions 8 - 10, Algorithm 7 finds an $\varepsilon$-optimal arm with probability at least $1 - \delta$.*

**Lemma 14.** *Using the notation of Algorithm 7:*

$$\mathbb{P}(|\hat{\mu}_0^1 - \mu_0^1| \geq \varepsilon/8 \vee |\hat{\mu}_0^4 - \mu_0^4| \geq \varepsilon/8) \leq \delta/2.$$

*Proof.* Based on the properties of subgaussian random variables:

$$\begin{aligned}
&\mathbb{P}(|\hat{\mu}_0^1 - \mu_0^1| \geq \varepsilon/8 \vee |\hat{\mu}_0^4 - \mu_0^4| \geq \varepsilon/8) \\
&\quad \leq \mathbb{P}(|\hat{\mu}_0^1 - \mu_0^1| \geq \varepsilon/8) + \mathbb{P}(|\hat{\mu}_0^4 - \mu_0^4| \geq \varepsilon/8) \\
&\quad \leq 2 \cdot 2 \cdot \exp \cdot \left(-n_0 \cdot (\varepsilon/8)^2/2\right) \\
&\quad \leq \delta_0 = \delta/2.
\end{aligned}$$

**Algorithm 7**

**Input:** $\varepsilon > 0$, $\delta > 0$

**Output:** an arm which is $\varepsilon$-optimal with probability at least $1 - \delta$

Set $\delta_0 = \delta/2$, $x_0^1 = 0$, $x_1^1 = 2^{m-2}$, $x_2^1 = 2^{m-1}$, $x_3^1 = 3 \cdot 2^{m-2}$, $x_4^1 = 2^m$.

Sample $n_0 = \lceil 128 \log(4/\delta_0)/\varepsilon^2 \rceil$ times $x_0^1$ and $x_4^1$.

Let $\hat{\mu}_i^\ell$ and $\mu_i^\ell$ denote the sample mean and the expectation of $x_i^\ell$.

Set $S_1 = \{i : 0 \le i \le 2^m\}$, $\delta_1 = \delta_0/2$, $\ell = 1$.

**while** $|S_\ell| > 5$ **do**

Sample the three arms: $x_1^\ell$, $x_2^\ell$, $x_3^\ell$, so that each of them will be sampled $n_\ell = \lceil 128 \log(6/\delta_\ell)/\varepsilon^2 \rceil$ times totally.

**if** $\hat{\mu}_1^\ell, \hat{\mu}_3^\ell \in (\hat{\mu}_2^\ell - \varepsilon/4, \hat{\mu}_2^\ell + \varepsilon/4)$ **then return** $x_2^\ell$

**else if** $\hat{\mu}_1^\ell \ge \hat{\mu}_2^\ell + \varepsilon/4$ **then** $S_{\ell+1} = \{i \in S_\ell : i \le x_2^\ell\}$

**else if** $\hat{\mu}_3^\ell \ge \hat{\mu}_2^\ell + \varepsilon/4$ **then** $S_{\ell+1} = \{i \in S_\ell : i \ge x_2^\ell\}$

**else if** $\hat{\mu}_1^\ell, \hat{\mu}_3^\ell \le \hat{\mu}_2^\ell - \varepsilon/4$ **then** $S_{\ell+1} = \{i \in S_\ell : x_1^\ell \le i \le x_3^\ell\}$

**else if** $\hat{\mu}_1^\ell \in (\hat{\mu}_2^\ell - \varepsilon/4, \hat{\mu}_2^\ell + \varepsilon/4)$ and $\hat{\mu}_3^\ell \le \hat{\mu}_2^\ell - \varepsilon/4$ **then**

Sample $n = \lceil 288 \log(6/\delta_\ell)/\varepsilon^2 \rceil - \lceil 128 \log(6/\delta_\ell)/\varepsilon^2 \rceil$ times $x_1^\ell$ and $x_2^\ell$ so that they are sampled $\lceil 288 \log(6/\delta_\ell)/\varepsilon^2 \rceil$ times totally.

Sample $n = \lceil 288 \log(12/\delta_\ell)/\varepsilon^2 \rceil$ times the arm $x_{1.5}^\ell = (x_1^\ell + x_2^\ell)/2$.

**if** $\hat{\mu}_1^\ell \le \hat{\mu}_{1.5}^\ell - \varepsilon/6$ or $\hat{\mu}_2^\ell \ge \hat{\mu}_{1.5}^\ell + \varepsilon/6$ **then** $S_{\ell+1} = \{i : x_1^\ell \le i \le x_3^\ell\}$

**else if** $\hat{\mu}_2^\ell \le \hat{\mu}_{1.5}^\ell - \varepsilon/6$ or $\hat{\mu}_1^\ell \ge \hat{\mu}_{1.5}^\ell + \varepsilon/6$ **then** $S_{\ell+1} = \{i : i \le x_2^\ell\}$

**else return** $x_{1.5}^\ell$ **end if**

**end if**

**else if** $\hat{\mu}_1^\ell \le \hat{\mu}_2^\ell - \varepsilon/4$ and $\hat{\mu}_3^\ell \in (\hat{\mu}_2^\ell - \varepsilon/4, \hat{\mu}_2^\ell + \varepsilon/4)$ **then**

Sample $n = \lceil 288 \log(6/\delta_\ell)/\varepsilon^2 \rceil - \lceil 128 \log(6/\delta_\ell)/\varepsilon^2 \rceil$ times $x_2^\ell$ and $x_3^\ell$ so that they are sampled $\lceil 288 \log(6/\delta_\ell)/\varepsilon^2 \rceil$ times totally.

Sample $n = \lceil 288 \log(12/\delta_\ell)/\varepsilon^2 \rceil$ times the arm $x_{2.5}^\ell = (x_2^\ell + x_3^\ell)/2$.

**if** $\hat{\mu}_2^\ell \le \hat{\mu}_{2.5}^\ell - \varepsilon/6$ or $\hat{\mu}_3^\ell \ge \hat{\mu}_{2.5}^\ell + \varepsilon/6$ **then** $S_{\ell+1} = \{i : i \le x_2^\ell\}$

**else if** $\hat{\mu}_3^\ell \le \hat{\mu}_{2.5}^\ell - \varepsilon/6$ or $\hat{\mu}_2^\ell \ge \hat{\mu}_{2.5}^\ell + \varepsilon/6$ **then**

$S_{\ell+1} = \{i \in S_\ell : x_1^\ell \le i \le x_3^\ell\}$

**else return** $x_{2.5}^\ell$ **end if**

**end if**

$x_0^{\ell+1} = \min S_{\ell+1}$, $x_4^{\ell+1} = \max S_{\ell+1}$, $x_1^{\ell+1} = (3 \cdot x_0^{\ell+1} + x_4^{\ell+1})/4$,

$x_2^{\ell+1} = (x_0^{\ell+1} + x_4^{\ell+1})/2$, $x_3^{\ell+1} = (x_0^{\ell+1} + 3 \cdot x_4^{\ell+1})/4$.

$\delta_{\ell+1} = \delta_\ell/2$, $\ell = \ell + 1$

**end while**

Sample the 5 arms so that each of them is sampled $n_\ell = \lceil 128 \log(6/\delta_\ell)/\varepsilon^2 \rceil$ times in total. Return the arm with the highest sample mean.

□

**Lemma 15.** *If $x$ is sampled $n_1 = \lceil 128 \log(6/\delta)/\varepsilon^2 \rceil$ times then*

$$\mathbb{P}(|\hat{\mu} - \mu_i| \geq \varepsilon/8) \leq \delta/3.$$

*Similarly, if $x$ is sampled $n_2 = \lceil 288 \log(6/\delta)/\varepsilon^2 \rceil$ times then*

$$\mathbb{P}(|\hat{\mu} - \mu_i| \geq \varepsilon/12) \leq \delta/3.$$

*Proof.* Based on Theorem 2:

$$\mathbb{P}(|\hat{\mu} - \mu_i| \geq \varepsilon/8) \leq 2 \exp \cdot \left(-n_1 \cdot (\varepsilon/8)^2/2\right) \leq \delta/3,$$

$$\mathbb{P}(|\hat{\mu} - \mu_i| \geq \varepsilon/12) \leq 2 \exp \cdot \left(-n_2 \cdot (\varepsilon/12)^2/2\right) \leq \delta/3.$$

□

**Lemma 16.** *Suppose that $|\hat{\mu}_i^\ell - \mu_i^\ell| \leq \varepsilon/8$ and $|\hat{\mu}_j^\ell - \mu_j^\ell| \leq \varepsilon/8$. Let $x^*$ denote the optimal arm. If $i < j$ and $\hat{\mu}_i^\ell \geq \hat{\mu}_j^\ell + \varepsilon/4$, then $x^* \leq x_j^\ell$. Similarly, if $i > j$ and $\hat{\mu}_i^\ell \geq \hat{\mu}_j^\ell + \varepsilon/4$, then $x^* \geq x_j^\ell$.*

*Proof.* If $i < j$ and $\hat{\mu}_i^\ell \geq \hat{\mu}_j^\ell + \varepsilon/4$, then $\mu_i^\ell \geq \mu_j^\ell$:

$$\mu_i^\ell \geq \hat{\mu}_i^\ell - \varepsilon/8 \geq \hat{\mu}_j^\ell + \varepsilon/4 - \varepsilon/8 = \hat{\mu}_j^\ell + \varepsilon/8 \geq \mu_j^\ell.$$

Arguing indirectly, assume that $x^* > x_j^\ell$. In this case, there exists a $t \in [0, 1]$ such that $x_j^\ell = t \cdot x_i^\ell + (1 - t) \cdot x^*$. Because of the concavity:

$$f(x_j^\ell) = f(t \cdot x_i^\ell + (1 - t) \cdot x^*) \geq t \cdot f(x_i^\ell) + (1 - t) \cdot f(x^*) > f(x_i^\ell).$$

This contradicts the fact that $\mu_i^\ell \geq \mu_j^\ell$.
Similarly, if $i > j$ and $\hat{\mu}_i^\ell \geq \hat{\mu}_j^\ell + \varepsilon/4$ then $\mu_i^\ell \geq \mu_j^\ell$. Indirectly assume that $x^* < x_j^\ell$. Then there exists a $t \in [0, 1]$ such that $x_j^\ell = t \cdot x_i^\ell + (1 - t) \cdot x^*$. Because of the concavity:

$$f(x_j^\ell) = f(t \cdot x_i^\ell + (1 - t) \cdot x^*) \geq t \cdot f(x_i^\ell) + (1 - t) \cdot f(x^*) > f(x_i^\ell).$$

This again contradicts the fact that $\mu_i^\ell \geq \mu_j^\ell$.

□

**Lemma 17.** *Suppose that* $|\hat{\mu}_i^\ell - \mu_i^\ell| \leq \varepsilon/12$, $|\hat{\mu}_j^\ell - \mu_j^\ell| \leq \varepsilon/12$. *Let* $x^*$ *denote the optimal arm. If* $i < j$ *and* $\hat{\mu}_i^\ell \geq \hat{\mu}_j^\ell + \varepsilon/6$, *then* $x^* \leq x_j^\ell$. *Similarly, if* $i > j$ *and* $\hat{\mu}_i^\ell \geq \hat{\mu}_j^\ell + \varepsilon/6$, *then* $x^* \geq x_j^\ell$.

*Proof.* It can be proved the same way as Lemma 16. □

**Lemma 18.** *If* $|\hat{\mu}_j^\ell - \mu_j^\ell| \leq \varepsilon/8$ *for* $j = 0, 1, 2, 3, 4$ *and* $\hat{\mu}_{i-1}^\ell, \hat{\mu}_{i+1}^\ell \in (\hat{\mu}_i^\ell - \varepsilon/4, \hat{\mu}_i^\ell + \varepsilon/4)$, *then*

$$\mu_{i-1}^\ell, \mu_{i+1}^\ell \in [\mu_i^\ell - \varepsilon/2, \mu_i^\ell + \varepsilon/2].$$

*Proof.* Based on the assumptions of the lemma:

$$\mu_{i-1}^\ell \geq \hat{\mu}_{i-1}^\ell - \varepsilon/8 \geq \hat{\mu}_i^\ell - \varepsilon/4 - \varepsilon/8 = \hat{\mu}_i^\ell - 3/8 \cdot \varepsilon \geq \mu_i^\ell - \varepsilon/8 - 3/8 \cdot \varepsilon = \mu_i^\ell - \varepsilon/2,$$

$$\mu_{i-1}^\ell \leq \hat{\mu}_{i-1}^\ell + \varepsilon/8 \leq \hat{\mu}_i^\ell + \varepsilon/4 + \varepsilon/8 = \hat{\mu}_i^\ell + 3/8 \cdot \varepsilon \leq \mu_i^\ell + \varepsilon/8 + 3/8 \cdot \varepsilon = \mu_i^\ell + \varepsilon/2.$$

Similarly, $\mu_{i+1}^\ell \in [\mu_i^\ell - \varepsilon/2, \mu_i^\ell + \varepsilon/2]$.

□

**Lemma 19.** *If* $|\hat{\mu}_j^\ell - \mu_j^\ell| \leq \frac{\varepsilon}{12}$ *for* $j = 0, 1, 2, 3, 4$ *and* $\hat{\mu}_{i-1}^\ell, \hat{\mu}_{i+1}^\ell \in (\hat{\mu}_i^\ell - \varepsilon/6, \hat{\mu}_i^\ell + \varepsilon/6)$, *then*

$$\mu_{i-1}^\ell, \mu_{i+1}^\ell \in [\mu_i^\ell - \varepsilon/3, \mu_i^\ell + \varepsilon/3].$$

*Proof.* Based on the assumptions of the lemma:

$$\mu_{i-1}^\ell \geq \hat{\mu}_{i-1}^\ell - \varepsilon/12 \geq \hat{\mu}_i^\ell - \varepsilon/6 - \varepsilon/12 \geq \mu_i^\ell - \varepsilon/12 - 3/12 \cdot \varepsilon = \mu_i^\ell - \varepsilon/3$$

$$\mu_{i-1}^\ell \leq \hat{\mu}_{i-1}^\ell + \varepsilon/12 \leq \hat{\mu}_i^\ell + \varepsilon/6 + \varepsilon/12 \leq \mu_i^\ell + \varepsilon/12 + 3/12 \cdot \varepsilon = \mu_i^\ell + \varepsilon/3.$$

Similarly, $\mu_{i+1}^\ell \in [\mu_i^\ell - \varepsilon/3, \mu_i^\ell + \varepsilon/3]$.

□

**Lemma 20.** *If* $|\hat{\mu}_i^\ell - \mu_i^\ell| \leq \varepsilon/8$ *for* $i = 1, 2, 3$ *and* $\hat{\mu}_1^\ell, \hat{\mu}_3^\ell \in (\hat{\mu}_2^\ell - \varepsilon/4, \hat{\mu}_2^\ell + \varepsilon/4)$, *then*

$$\mu_2^\ell \geq \max_{x \in S_\ell} f(x) - \varepsilon.$$

*Proof.* Let $x^* = \arg\max_{x \in S_\ell} f(x)$. Arguing indirectly assume, that $f(x^*) > f(x_2^\ell) + \varepsilon$.

If $x^* < x_1^\ell$, then there exists a $t \in [0,1]$ such that $x_1^\ell = (1-t) \cdot x^* + t \cdot x_2^\ell$. As $x_1^\ell$ is closer to $x^*$ than to $x_2^\ell$, $t \leq 1/2$. Because of the concavity:

$$\begin{aligned} f(x_1^\ell) &= f((1-t) \cdot x^* + t \cdot x_2^\ell) \\ &\geq (1-t) \cdot f(x^*) + t \cdot f(x_2^\ell) \\ &> (1-t) \cdot (f(x_2^\ell) + \varepsilon) + t \cdot f(x_2^\ell) \\ &= f(x_2^\ell) + (1-t) \cdot \varepsilon \\ &\geq f(x_2^\ell) + \varepsilon/2. \end{aligned}$$

This contradicts the fact that $f(x_1^\ell) \leq f(x_2^\ell) + \varepsilon/2$ based on Lemma 18.
If $x_1^\ell < x^* < x_2^\ell$, then there exists a $t \in [0,1]$ such that $x_2^\ell = (1-t) \cdot x^* + t \cdot x_3^\ell$. As $x_2^\ell$ is closer to $x^*$ than to $x_3^\ell$, $t \leq 1/2$. Because of the concavity:

$$\begin{aligned} f(x_2^\ell) &= f((1-t) \cdot x^* + t \cdot x_3^\ell) \\ &\geq (1-t) \cdot f(x^*) + t \cdot f(x_3^\ell) \\ &> (1-t) \cdot (f(x_2^\ell) + \varepsilon) + t \cdot (f(x_2^\ell) - \varepsilon/2) \\ &= f(x_2^\ell) + (1 - 3/2 \cdot t) \cdot \varepsilon \\ &> f(x_2^\ell). \end{aligned}$$

This is again a contradiction.
If $x_2^\ell < x^* < x_3^\ell$, then there exists a $t \in [0,1]$ such that $x_2^\ell = (1-t) \cdot x_1^\ell + t \cdot x^*$. As $x_2^\ell$ is closer to $x^*$ than to $x_1^\ell$, $1 > t \geq 1/2$. Because of the concavity:

$$\begin{aligned} f(x_2^\ell) &= f((1-t) \cdot x_1^\ell + t \cdot x^*) \\ &\geq (1-t) \cdot f(x_1^\ell) + t \cdot f(x^*) \\ &> (1-t) \cdot (f(x_2^\ell) - \varepsilon/2) + t \cdot (f(x_2^\ell) + \varepsilon) \\ &= f(x_2^\ell) + (3/2 \cdot t - 1/2) \cdot \varepsilon \\ &> f(x_2^\ell). \end{aligned}$$

This way we got a contradiction.
If $x_3^\ell < x^*$, then there exists a $t \in [0,1]$ such that $x_3^\ell = (1-t) \cdot x_2^\ell + t \cdot x^*$.As $x_3^\ell$ is

closer to $x^*$ than to $x_2^\ell$, $t \geq 1/2$. Because of the concavity:

$$\begin{aligned} f(x_3^\ell) &= f((1-t)\cdot x_2^\ell + t\cdot x^*) \\ &\geq (1-t)\cdot f(x_2^\ell) + t\cdot f(x^*) \\ &> (1-t)\cdot f(x_2^\ell) + t\cdot (f(x_2^\ell)+\varepsilon) \\ &= f(x_2^\ell) + t\cdot\varepsilon \\ &\geq f(x_2^\ell) + \varepsilon/2. \end{aligned}$$

This contradicts the fact that $f(x_3^\ell) \leq f(x_2^\ell) + \varepsilon/2$ based on Lemma 18.

□

**Lemma 21.** *If $|\hat{\mu}_i^\ell - \mu_i^\ell| \leq \varepsilon/12$ for $i = 1, 1.5, 2$ and $\hat{\mu}_1^\ell, \hat{\mu}_2^\ell \in (\hat{\mu}_{1.5}^\ell - \varepsilon/6, \hat{\mu}_{1.5}^\ell + \varepsilon/6)$, then*

$$\mu_{1.5}^\ell \geq \max_{x\in\{i: x_0^\ell \leq i \leq x_3^\ell\}} f(x) - \varepsilon.$$

*Similarly, if $|\hat{\mu}_i^\ell - \mu_i^\ell| \leq \varepsilon/12$ for $i = 2, 2.5, 3$ and $\hat{\mu}_2^\ell, \hat{\mu}_3^\ell \in (\hat{\mu}_{2.5}^\ell - \varepsilon/6, \hat{\mu}_{2.5}^\ell + \varepsilon/6)$, then*

$$\mu_{2.5}^\ell \geq \max_{x\in\{i: x_1^\ell \leq i \leq x_4^\ell\}} f(x) - \varepsilon.$$

*Proof.* Let's introduce the following notation

$$x^* = \operatorname*{arg\,max}_{x\in\{i: x_0^\ell \leq i \leq x_3^\ell\}} f(x).$$

Arguing indirectly assume, that $f(x^*) > f(x_{1.5}^\ell) + \varepsilon$. If $x^* < x_1^\ell$, then there exists a $t \in [0,1]$ such that $x_1^\ell = (1-t)\cdot x^* + t\cdot x_{1.5}^\ell$. From this $t = \frac{x_1 - x^*}{x_{1.5} - x^*} \leq \frac{2}{3}$. Because of the concavity:

$$\begin{aligned} f(x_1^\ell) &= f((1-t)\cdot x^* + t\cdot x_{1.5}^\ell) \\ &\geq (1-t)\cdot f(x^*) + t\cdot f(x_{1.5}^\ell) \\ &> (1-t)\cdot (f(x_{1.5}^\ell)+\varepsilon) + t\cdot f(x_{1.5}^\ell) \\ &= f(x_{1.5}^\ell) + (1-t)\cdot\varepsilon \\ &\geq f(x_{1.5}^\ell) + \varepsilon/3. \end{aligned}$$

This contradicts the fact that $f(x_1^\ell) \leq f(x_{1.5}^\ell) + \varepsilon/3$ based on Lemma 19.

If $x_1^\ell < x^* < x_{1.5}^\ell$, then there exists a $t \in [0,1]$ such that $x_{1.5}^\ell = (1-t) \cdot x^* + t \cdot x_2^\ell$. Here $t = \frac{x_{1.5} - x^*}{x_2 - x^*} \leq \frac{1}{2}$. Because of the concavity:

$$\begin{aligned} f(x_{1.5}^\ell) &= f((1-t) \cdot x^* + t \cdot x_2^\ell) \\ &\geq (1-t) \cdot f(x^*) + t \cdot f(x_2^\ell) \\ &> (1-t) \cdot (f(x_{1.5}^\ell) + \varepsilon) + t \cdot (f(x_{1.5}^\ell) - \varepsilon/3) \\ &= f(x_{1.5}^\ell) + (1 - 4/3 \cdot t) \cdot \varepsilon \\ &> f(x_{1.5}^\ell). \end{aligned}$$

This is again a contradiction.
If $x_{1.5}^\ell < x^* < x_2^\ell$, then there exists a $t \in [0,1]$ such that $x_{1.5}^\ell = (1-t) \cdot x_1^\ell + t \cdot x^*$. In this case $t = \frac{x_{1.5} - x_1}{x^* - x_1} \geq \frac{1}{2}$. Because of the concavity:

$$\begin{aligned} f(x_{1.5}^\ell) &= f((1-t) \cdot x_1^\ell + t \cdot x^*) \\ &\geq (1-t) \cdot f(x_1^\ell) + t \cdot f(x^*) \\ &> (1-t) \cdot (f(x_{1.5}^\ell) - \varepsilon/3) + t \cdot (f(x_{1.5}^\ell) + \varepsilon) \\ &= f(x_{1.5}^\ell) + (4/3 \cdot t - 1/3) \cdot \varepsilon \\ &> f(x_{1.5}^\ell). \end{aligned}$$

This way we got another contradiction.
If $x_2^\ell < x^*$, then there exists a $t \in [0,1]$ such that $x_2^\ell = (1-t) \cdot x_{1.5}^\ell + t \cdot x^*$. In this case $t = \frac{x_2 - x_{1.5}}{x^* - x_{1.5}} \geq \frac{1}{3}$. Because of the concavity:

$$\begin{aligned} f(x_2^\ell) &= f((1-t) \cdot x_{1.5}^\ell + t \cdot x^*) \\ &\geq (1-t) \cdot f(x_{1.5}^\ell) + t \cdot f(x^*) \\ &> (1-t) \cdot f(x_{1.5}^\ell) + t \cdot (f(x_{1.5}^\ell) + \varepsilon) \\ &= f(x_{1.5}^\ell) + t \cdot \varepsilon \\ &\geq f(x_{1.5}^\ell) + \varepsilon/3. \end{aligned}$$

This contradicts the fact that $f(x_2^\ell) \leq f(x_{1.5}^\ell) + \varepsilon/3$ based on Lemma 19. □

**Lemma 22.** *If $|S_\ell| = 5$ and $|\hat{\mu_j} - \mu_j| \leq \varepsilon/8$ for $j = 1,2,3,4,5$ then the arm with the highest sample mean is $\varepsilon$-optimal.*

*Proof.* Suppose, that $\hat{\mu}_i$ is the highest sample mean. In this case for $j = 1, 2, 3, 4, 5$:

$$\mu_i \geq \hat{\mu}_i - \varepsilon/8 \geq \hat{\mu}_j - \varepsilon/8 \geq \mu_j - \varepsilon/4.$$

□

*Proof of Theorem.* Based on the lemmas, the probability that there is a round in which for a sampled arm $|\hat{\mu} - \mu| \geq \varepsilon/8$ (or $|\hat{\mu} - \mu| \geq \varepsilon/12$ when needed) is less than $\delta$. Now consider the case, when $|\hat{\mu} - \mu| \leq \varepsilon/8$ (or $|\hat{\mu} - \mu| \leq \varepsilon/12$ when needed) for all of the arms sampled in any of the rounds. In each round the set of the arms is halved or the algorithm terminates. By Lemma 16 and 17 the optimal arm is never removed and by Lemma 20, 21 and 22 an $\varepsilon$-optimal arm is returned when the algorithm terminates. This way Algorithm 7 returns an $\varepsilon$-optimal arm with probability at least $1 - \delta$. □

**Theorem 12.** *The sample complexity of Algorithm 7 in case of $n = 2^m + 1$ arms is:*

$$\mathcal{O}\left(\frac{1}{\varepsilon^2}\left(m^2 + m\log\frac{1}{\delta}\right)\right).$$

*Proof.* At the beginning we sample two arms $n_0 = \lceil 128\log(4/\delta_0)/\varepsilon^2 \rceil$ times. If in a round $x_{1.5}$ or $x_{2.5}$ is sampled, then we will count these samples to the next round. We can do this, because otherwise this arm would have been sampled in the next round, but this way it is not needed anymore. This way in each round three arms are sampled, each $n = \lceil 288\log(6/\delta_\ell)/\varepsilon^2 \rceil$ times max. When only five arms remain, then three of them is already sampled, so in the last round only two arms have to

be sampled, $n = \lceil 288 \log(6/\delta_{m-1})/\varepsilon^2 \rceil$ times maximum. So the sample complexity:

$$\begin{aligned}
&2 \cdot \lceil 128 \log(4/\delta_0)/\varepsilon^2 \rceil + 3 \cdot \sum_{i=1}^{m-2} \lceil 288 \log(6/\delta_\ell)/\varepsilon^2 \rceil \\
&\quad + 2 \cdot \lceil 288 \log(6/\delta_{m-1})/\varepsilon^2 \rceil \\
&\quad \leq 3m + 3 \cdot \frac{288}{\varepsilon^2} \sum_{i=1}^{m-1} \log \frac{6}{\delta_i} \\
&\quad = 3m + 3 \cdot \frac{288}{\varepsilon^2} \log \left( \prod_{i=1}^{m-1} \frac{12 \cdot 2^i}{\delta} \right) \\
&\quad = 3m + 3 \cdot \frac{288}{\varepsilon^2} \log \left( \frac{12^{m-1} \cdot 2^{(m-1)m/2}}{\delta^{m-1}} \right) \\
&\quad = 3m + 3 \cdot \frac{288}{\varepsilon^2} \left( (m-1) \log \frac{12}{\delta} + \frac{(m-1)m}{2} \log 2 \right) \\
&\quad = \mathcal{O}\left( \frac{1}{\varepsilon^2} \left( m^2 + m \log \frac{1}{\delta} \right) \right).
\end{aligned}$$

□

If there are $2^m + 1 < n < 2^{m+1} + 1$ arms, we can do the following:
Run the first round of Algorithm 7 with arms

$$\begin{aligned}
&\left\{ i : \left\lfloor \frac{n}{2} \right\rfloor - 2^{m-1} \leq i \leq \left\lfloor \frac{n}{2} \right\rfloor + 2^{m-1} \right\}, \\
&x_0^1 = \left\lfloor \frac{n}{2} \right\rfloor - 2^{m-1},\ x_1^1 = \left\lfloor \frac{n}{2} \right\rfloor - 2^{m-2},\ x_2^1 = \left\lfloor \frac{n}{2} \right\rfloor, \\
&x_3^1 = \left\lfloor \frac{n}{2} \right\rfloor + 2^{m-2},\ x_4^1 = \left\lfloor \frac{n}{2} \right\rfloor + 2^{m-1}
\end{aligned}$$

and $\delta_0 = \delta/2$. If after the first round of Algorithm 7:

- $S_2 = \{i : x_0^1 \leq i \leq x_2^1\}$, then set $S = \{i : 0 \leq i \leq 2^m\}$,
- $S_2 = \{i : x_1^1 \leq i \leq x_3^1\}$, then set $S = \left\{ i : \left\lfloor \frac{n}{2} \right\rfloor - 2^{m-1} \leq i \leq \left\lfloor \frac{n}{2} \right\rfloor + 2^{m-1} \right\}$,
- $S_2 = \{i : x_2^1 \leq i \leq x_4^1\}$, then set $S = \{i : n - 2^m \leq i \leq n\}$.

The length of $S$ is $2^m + 1$ in all cases. Now do Algorithm 7 with $S$ and $\delta_0 = \delta/8$. Based on the lemmas, the probability that the optimal arm is discarded in the first round is less than $\frac{3}{4} \cdot \delta$. If the optimal arm is not removed in the first round, then

Algorithm 7 with $S$ and $\delta_0 = \delta/8$ will return an $\varepsilon$-optimal arm with probability at least $1 - \delta/4$. So the probability that the returned arm is not $\varepsilon$-optimal is less than $\delta$.

At the beginning two arms are sampled $\lceil 128 \log(4/\delta_0)/\varepsilon^2 \rceil$ times each, then we sample three arms $\lceil 288 \log(6/\delta_\ell)/\varepsilon^2 \rceil$ times maximum. After that we do Algorithm 7 with $\delta_0 = \delta/8$ with $2^m + 1$ arms, so based on our previous calculation, the sample complexity in this case is:

$$\begin{aligned} &2 \cdot \lceil 128 \log(4/\delta_0)/\varepsilon^2 \rceil + 3 \cdot \lceil 288 \log(6/\delta_\ell)/\varepsilon^2 \rceil \\ &\quad + \mathcal{O}\left(\frac{1}{\varepsilon^2}\left(m^2 + m \log \frac{1}{\delta}\right)\right) \\ &\quad = \mathcal{O}\left(\frac{1}{\varepsilon^2}\left(m^2 + m \log \frac{1}{\delta}\right)\right). \end{aligned}$$

Similarly to the Lipschitz continuous case, if there is a known $\Delta > 0$ such that $|\mu_i - \mu_{i-1}| \leq \Delta$, $i = 1, 2, ..., n$, then Algorithm 7 can terminate when the number of arms is $2 \cdot \lfloor \frac{\varepsilon}{\Delta} \rfloor + 1$ or less, by returning the arm in the middle ($x_j$). In this case the returned arm will be $\varepsilon$-optimal because:

$$|\mu_i - \mu_j| \leq |i - j| \cdot \Delta \leq \left\lfloor \frac{\varepsilon}{\Delta} \right\rfloor \cdot \Delta \leq \varepsilon \quad \forall i.$$

In this case the algorithm can terminate after $\left\lfloor \log_2 \frac{n}{2\lfloor \varepsilon/\Delta \rfloor + 1} \right\rfloor$ rounds so the sample complexity in this case is:

$$\mathcal{O}\left(\frac{\ell^2}{\varepsilon^2} + \frac{\ell}{\varepsilon^2} \log \frac{1}{\delta}\right).$$

where $\ell = \left\lfloor \log_2 \frac{n}{2\lfloor \varepsilon/\Delta \rfloor + 1} \right\rfloor$.

## 8.2 Experiments

Based on Theorem 12, the sample complexity of the algorithm is

$$\mathcal{O}\left(\frac{1}{\varepsilon^2}\left(m^2 + m \log \frac{1}{\delta}\right)\right).$$

The following plots show the relationship between the sample complexity of Algorithm 7 and the parameters of the algorithm, when the $f(x) = -(x - 0.5 \cdot N)^2$

function describes the expectation of the arms in case of $N$ arms. Figure 13 shows the relationship between the sample complexity and the $\varepsilon$ parameter. We can observe, that the sample complexity decreases steeply as the value of $\varepsilon$ increases, which aligns with the formula of sample complexity. Figure 14 explores the dependence of sample complexity on $\delta$, and reveals the inverse relationship between them, which is consistent with the $\log \frac{1}{\delta}$ term in the formula of sample complexity.

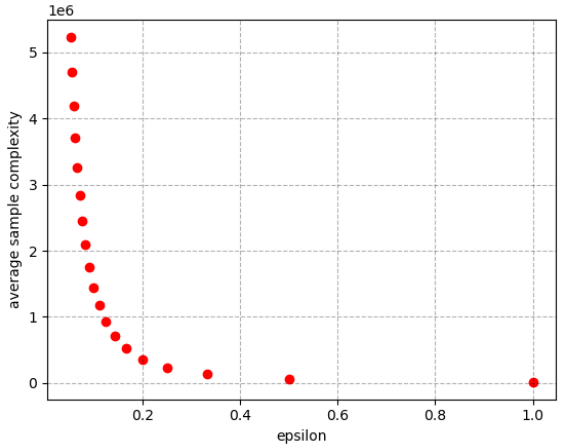



Figure 13: The relationship between the sample complexity and $\varepsilon$ in case of $\delta = 0.05$ and $N = 33$

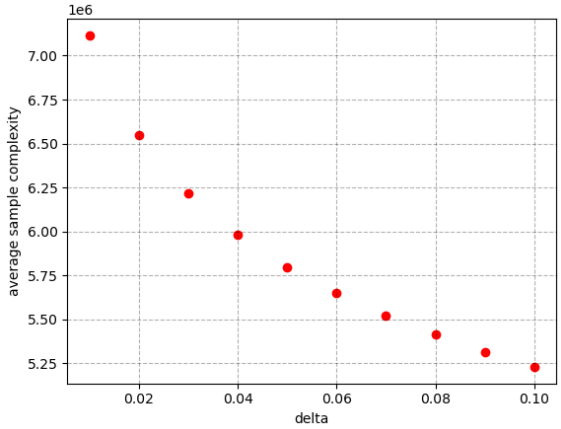



Figure 14: The relationship between the sample complexity and $\delta$ in case of $\varepsilon = 0.05$ and $N = 33$

Figure 15 and 16 show the relationship between the sample complexity and the number of arms in case of two different values of $\delta$. For larger $\delta$ values, the relationship appears to be quadratic, while for smaller values, it seems more linear. This is consistent with the fact that for higher values of $\delta$, the $m^2$ term dominates in the formula of the sample complexity, while for smaller $\delta$ values the $m \log \frac{1}{\delta}$ term becomes more significant.

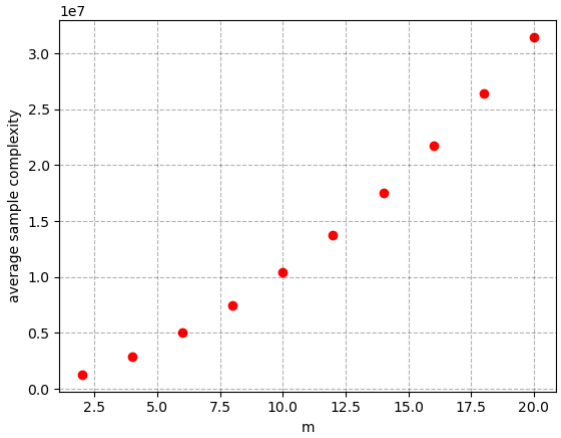



Figure 15: The relationship between the sample complexity and the number of arms ($N = 2^m + 1$) in case of $\varepsilon = 0.05$ and $\delta = 0.5$

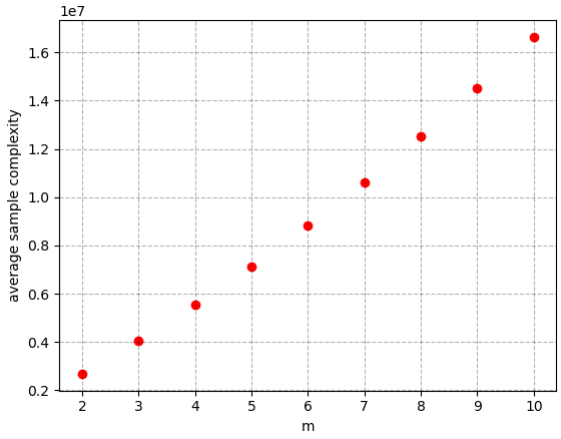



Figure 16: The relationship between the sample complexity and the number of arms ($N = 2^m + 1$) in case of $\varepsilon = 0.05$ and $\delta = 0.01$

# 9 Conclusion

## 9.1 Overview of Known Results

In my thesis, I presented some of the principal results related to stochastic multi-armed bandit problems. I explored two fundamental algorithms, each addressing a different objective.

First, I presented the Upper Confidence Bound algorithm, which is one of the best-known algorithms for maximizing the cumulative reward. Under this setting, the regret is an important measure of performance, since it measures the difference between the maximal reward, that can be achieve by playing the optimal arm all the time and the expectation of the sum of the rewards collected by the algorithm. By analyzing the regret of the algorithm, I demonstrated the effectiveness of the algorithm in balancing exploration and exploitation to maximize the reward in the long run.

Next, I examined the Median Elimination Algorithm, which can be used to identify close-to-optimal arms in the multi-armed bandit problem. I proved, that the algorithm returns a near-optimal arm with high probability and calculated the sample complexity of the algorithm. I have also provided a matching lower bound on the expected number of trials under any $(\varepsilon, \delta) - PAC$ policy.

## 9.2 Contributions of the Thesis

I have also presented my own results regarding some special cases of the multi-armed bandit problem, in which further information is known about the expectations of the arms. In all of these cases I assumed that the arms are 1-subgaussian. The main contributions of this thesis are the following:

- I have developed an $(\varepsilon, \delta) - PAC$ algorithm for the case, when there are finitely many arms, and we know that the expectations of the arms increase up to a certain arm, after which the expectation decreases, and the difference between the expectations of neighboring arms is at least $\Delta$, where $\Delta$ is known in advance. In this case the optimal arm could be found using the

Median Elimination algorithm (by choosing $\varepsilon < \Delta$), however the provided new algorithm finds the optimal arm much faster, with a sample complexity of

$$\mathcal{O}\left(\log n + \frac{1}{\Delta^2}\log\frac{n}{\delta}\right)$$

instead of

$$\mathcal{O}\left(\frac{n}{\Delta^2}\log\frac{1}{\delta}\right).$$

- I have also considered the case, when the arms are the points of the $[0,1]$ interval and an unknown concave function describes the expectations of the arms and developed an $(\varepsilon,\delta) - PAC$ algorithm for this case as well. This algorithm can be improved further, if we now that the function is Lipschitz continuous with a known Lipschitz contstant $L$. In this case, the algorithm can terminate, if the interval of the remaining arms is small enough, since in this case we can estimate the difference in the expectations of the arms in the interval, using the fact, that the function is Lipschitz continuous. I have also calculated the sample complexity of the algorithm.

- After this, I have studied the the discrete case, when an unknown concave function describes the expectations of the arms. I have modified the algorithm created for the continuous case to work in the discrete case as well. The algorithm in the continuous case removed at least one quarter of the arms in each step. The idea was to modify the algorithm, so that it removes half of the arms in each round. This way, the algorithm can work, when there are $2^m + 1$ arms for some $m \in \mathbb{N}$, since we can always sample the arm exactly in the middle of two arms, this was obvious in the continuous case, but not in the discrete case. In this case the Median Elimination algorithm could be used to find an $\varepsilon$-optimal arm with probability at least $1-\delta$ with a sample complexity of

  $$\mathcal{O}\left(\frac{n}{\varepsilon^2}\log\frac{1}{\delta}\right).$$

  However, the algorithm I developed achieves the same result with a sample

complexity of

$$\mathcal{O}\left(\frac{1}{\varepsilon^2}\left((\log n)^2 + \log n \cdot \log\frac{1}{\delta}\right)\right).$$

- I have implemented each of these algorithms and run experiments to visualize the relationship between he algorithm inputs and the sample complexity.

## 9.3 Future Directions

In this thesis, we focused on cases of the stochastic multi-armed bandit problem, where the expectations of the arms are described by an unknown concave function. However, the presented algorithms assume no additional external information.

A promising direction for future research is extending this framework to contextual bandits, where the distributions of the rewards depend not only on the selected arm but also on observable contextual information. Such an extension is important, because in many real-world applications additional data can significantly enhance decision-making. For example, recommendation systems can use demographic information to provide better suggestions.

By incorporating a concave reward function in the contextual bandit framework, it might be possible to develop algorithms that can leverage both contextual information and the concave structure of the reward function to outperform the existing algorithms.

# A Inequalities in Probability Theory

**Statement 3.** *(Markov's inequality)*
*For any random variable $X$ and $\varepsilon > 0$:*

$$\mathbb{P}(|X| \geq \varepsilon) \leq \frac{\mathbb{E}[|X|]}{\varepsilon}.$$

**Statement 4.** *(Doob's maximal inequality)*
*Let $(X_t)_{t=0}^n$ be a submartingle with $X_t \geq 0$ almost surely for all $t$. Then for any $\varepsilon > 0$,*

$$\mathbb{P}\left(\max_{0 \leq t \leq n} X_t \geq \varepsilon\right) \leq \frac{\mathbb{E}[X_n]}{\varepsilon}.$$

# References


[1] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, second edition, 2018.

[2] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[3] Alexandra Iacob, Bogdan Cautis, and Silviu Maniu. Contextual bandits for advertising campaigns: A diffusion-model independent approach. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 513–521. SIAM, 2022.

[4] Vashist Avadhanula, Riccardo Colini Baldeschi, Stefano Leonardi, Karthik Abinav Sankararaman, and Okke Schrijvers. Stochastic bandits for multi-platform budget optimization in online advertising. In *Proceedings of the Web Conference 2021*, pages 2805–2817, 2021.

[5] Hongbo Guo, Ruben Naeff, Alex Nikulkov, and Zheqing Zhu. Evaluating online bandit exploration in large-scale recommender system. In *KDD-23 Workshop on Multi-Armed Bandits and Reinforcement Learning: Advancing Decision Making in E-Commerce and Beyond*, 2023.

[6] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, WWW '10, page 661–670. ACM, April 2010.

[7] Liu Leqi, Giulio Zhou, Fatma Kilinc-Karzan, Zachary Lipton, and Alan Montgomery. A field test of bandit algorithms for recommendations: Understanding the validity of assumptions on human preferences in multi-armed bandits. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, page 1–16. ACM, April 2023.

[8] Yogatheesan Varatharajah and Brent Berry. A contextual-bandit-based approach for informed decision-making in clinical trials. *Life*, 12(8):1277, 2022.

[9] Sofía S. Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statistical Science*, 30(2), May 2015.

[10] Matteo Gagliolo and Jürgen Schmidhuber. *Algorithm Selection as a Bandit Problem with Unbounded Losses*, page 82–96. Springer Berlin Heidelberg, 2010.

[11] Mengying Zhu, Xiaolin Zheng, Yan Wang, Yuyuan Li, and Qianqiao Liang. Adaptive portfolio by solving multi-armed bandit via thompson sampling. *arXiv preprint arXiv:1911.05309*, 2019.

[12] Jonas W Mueller, Vasilis Syrgkanis, and Matt Taddy. Low-rank bandit methods for high-dimensional dynamic pricing. *Advances in Neural Information Processing Systems*, 32, 2019.

[13] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(39):1079–1105, 2006.

[14] Eyal Even-Dar, Shie Mannor, and Y. Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *Annual Conference Computational Learning Theory*, 2002.

[15] Shie Mannor and John N Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648, 2004.

[16] Balázs Csanád Csáji and Erik Weyer. System identification with binary observations by stochastic approximation and active learning. In *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pages 3634–3639. IEEE, 2011.