Stochastic Recursive Optimization: A Structured Multi-Armed Bandit Problem

Roland Szögi
Supervisor: Balázs Csanád Csáji

I. Introduction

A multi-armed bandit problem is a problem in which a series of decisions have to be made in order to maximize the expected reward while having partial knowledge of the usefulness of the actions. However, by choosing an action, we get information about the usefulness of that specific action. The multi-armed bandit problem is one of the most studied problems in decision theory [1] with many applications including A/B testing, advert placement and recommendation services [2].

I have investigated a special case of the multi-armed bandit problem, in which further information about the structure of the arms is known. I have developed an algorithm that returns an arm which is close to optimal with high probability.

II. Multi-armed Bandits

The multi-armed bandit model consists of a set of arms $\mathcal{A}$ ($n = |\mathcal{A}|$) and to every arm $a \in \mathcal{A}$ belongs a distribution $\nu(a)$. In each round an arm $a \in \mathcal{A}$ is chosen and a reward $R(a)$ is sampled from distribution $\nu(a)$. 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 1. *An arm $a \in \mathcal{A}$ is called ε-optimal if*

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

One of the most common learning objectives is to find an ε-optimal arm with high probability.

Definition 2. *An algorithm is called an (ε, δ)-PAC (probably approximately correct) algorithm for the multi-armed bandit problem with sample complexity T, if it outputs an ε-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.*

An (ε, δ)-PAC algorithm is known for the case of binary rewards, called Median Elimination [3].

Statement 1. *The Median Elimination algorithm is an (ε, δ)-PAC algorithm and its sample complexity is*

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

In [4] an $\mathcal{O}\left((n/\varepsilon^2)\log(1/\delta)\right)$ lower bound is provided on the expected number of trials under any policy that finds an ε-optimal arm with probability at least $1 - \delta$.

Algorithm 1 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 $1/(\varepsilon_\ell/2)^2 \log(3/\delta_\ell)$ times, and let $\hat{p}_a^\ell$ denote its empirical value
    Find the median of $\hat{p}_a^\ell$, denoted by $m_\ell$
    $S_{\ell+1} = S_\ell \setminus \{a : \hat{p}_a^\ell < m_\ell\}$
    $\varepsilon_{\ell+1} = \frac{3}{4}\varepsilon_\ell$, $\delta_{\ell+1} = \delta_\ell/2$, $\ell = \ell + 1$
  until $|S_\ell| = 1$
```

III. Subgaussian Random Variables

More information about subgaussian random variables and the proof of Statement 2 can be found in [2].

Definition 3. *A random variable X is σ-subgaussian if for all $\lambda \in \mathbb{R}$:*

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

Statement 2. *Assume that $X_i - \mu$ are independent, σ-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 $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$.

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

IV. Arms with a Special Structure

In the previous semester I have investigated a special case of the multi-armed bandit problem, in which we have further knowledge of the structure of the arms. This special case of the multi-armed bandit problem also arises in practice, for example the quantized estimation problem studied in [5] 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 numbered from 0 to 2^m: $a_0, a_1, ..., a_{2^m}$.*
(The expectation of arm a_i will be denoted by μ_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 $\Delta > 0$ such that*

$$|\mu_{i+1} - \mu_i| \geq \Delta \quad \forall\, i \in \{0, 1, ..., 2^m - 1\},$$

and it is known in advance.

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

Algorithm 2

```
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 the three arms: $a_j$ , $j \in \{i \cdot 2^{m-\ell-1}$, $i = 1, 2, 3\}$ $n_\ell = \lceil \log(4/\delta_\ell)/(2^{2m-2\ell-5}\Delta^2) \rceil$ times each, and let $\hat{\mu}_j^\ell$ denote their empirical values
    $i_\ell^* = \arg\max_j \hat{\mu}_j^\ell$
    $S_{\ell+1} = \left\{a_i : i_\ell^* - 2^{m-\ell-1} \leq i \leq i_\ell^* + 2^{m-\ell-1}\right\}$
    Renumber the arms from 0 to $2^{m-\ell}$
    $\delta_{\ell+1} = \delta_\ell/2$, $\ell = \ell + 1$
  end while
  Sample each of the three remaining arms $a_j$, $j \in \{0, 1, 2\}$ $n_m = \lceil \log(4/\delta_m)/(2^{-3}\Delta^2) \rceil$ times, and let $\hat{\mu}_j^m$ denote their empirical values
  $i_m^* = \arg\max_j \hat{\mu}_j^m$
  return $a_{i_m^*}$
```

Theorem 1. *Under Assumptions 1-4, Algorithm 2 finds the optimal arm with probability at least $1 - \delta$ and its sample complexity is*

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

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 $a_{i_1^*}$, the set of these arms will be S_2. Renumber the arms from 0 to 2^{m-1}. Set $\delta_2 = \delta/4$ and $\ell = 2$. After that we can continue with the second round of Algorithm 2.

V. Infinitely Many Arms with a Concave Structure in 1D

In this semester I have considered the problem in which the arms are the elements of the $[0, 1]$ interval and an unknown concave function describes the expectations of the arms. The assumptions are the following:

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

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

Algorithm 3

```
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 the two arms: $x_0^1$ and $x_4^1$. Let $\hat{\mu}_0^1$ and $\hat{\mu}_4^1$ denote the sample means and $\mu_0^1$, $\mu_4^1$ denote the corresponding 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 $\hat{\mu}_1^\ell, \hat{\mu}_2^\ell, \hat{\mu}_3^\ell$ denote the sample means and $\mu_1^\ell, \mu_2^\ell, \mu_3^\ell$ denote the 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 $x_2^\ell$
    end if
    if $\hat{\mu}_1^\ell \geq \hat{\mu}_2^\ell + \varepsilon/4$ then
      $S_{\ell+1} = S_{\ell+1} \setminus (x_2^\ell, x_4^\ell]$
    else if $\hat{\mu}_1^\ell \leq \hat{\mu}_2^\ell - \varepsilon/4$ then
      $S_{\ell+1} = S_{\ell+1} \setminus [x_0^\ell, x_1^\ell)$
    end if
    if $\hat{\mu}_3^\ell \geq \hat{\mu}_2^\ell + \varepsilon/4$ then
      $S_{\ell+1} = S_{\ell+1} \setminus [x_0^\ell, x_2^\ell)$
    else if $\hat{\mu}_3^\ell \leq \hat{\mu}_2^\ell - \varepsilon/4$ then
      $S_{\ell+1} = S_{\ell+1} \setminus (x_3^\ell, x_4^\ell]$
    end if
    $x_0^{\ell+1} = \min S_{\ell+1}$, $x_4^{\ell+1} = \max S_{\ell+1}$,
    $x_1^{\ell+1} = \frac{3}{4} \cdot x_0^{\ell+1} + \frac{1}{4} \cdot x_4^{\ell+1}$,
    $x_2^{\ell+1} = \frac{1}{2} \cdot x_0^{\ell+1} + \frac{1}{2} \cdot x_4^{\ell+1}$,
    $x_3^{\ell+1} = \frac{1}{4} \cdot x_0^{\ell+1} + \frac{3}{4} \cdot x_4^{\ell+1}$.
    Let $\hat{\mu}_0^{\ell+1}$ and $\hat{\mu}_4^{\ell+1}$ 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 the expectations of $x_0^{\ell+1}$ and $x_4^{\ell+1}$.
    $\delta_{\ell+1} = \delta_\ell/2$, $\ell = \ell + 1$
  end while
```

In each round of Algorithm 3 the set of arms is reduced or the algorithm terminates. In round ℓ the set of arms is denoted by S_ℓ, which is divided into four subsets with equal size by five arms: x_0^ℓ, x_1^ℓ, x_2^ℓ, x_3^ℓ, x_4^ℓ. The expectation of arm x_i^ℓ is denoted by μ_i^ℓ. In round ℓ we have estimation of μ_0^ℓ and μ_4^ℓ from the previous round. We want to estimate μ_1^ℓ, μ_2^ℓ and μ_3^ℓ as well, so we sample x_1^ℓ, x_2^ℓ, x_3^ℓ many times and we estimate the expectations with the sample means. The estimation of the expectation of arm x_i^ℓ is denoted by $\hat{\mu}_i^\ell$. If $\hat{\mu}_1^\ell$ and $\hat{\mu}_3^\ell$ are close to $\hat{\mu}_2^\ell$, then the arm x_2^ℓ will be close-to-optimal because of the concavity and the algorithm returns x_2^ℓ. If $\hat{\mu}_1^\ell$ is much bigger 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^ℓ and we can remove these arms. If $\hat{\mu}_1^\ell$ is much smaller than $\hat{\mu}_2^\ell$, then $\mu_1^\ell \leq \mu_2^\ell$ and because of the concavity, the optimal arm cannot be on the right of arm x_1^ℓ and we can remove these arms. Similarly, if $\hat{\mu}_3^\ell$ is much smaller or bigger than $\hat{\mu}_2^\ell$, we can also remove arms.

Theorem 2. *Under Assumptions 5-6, Algorithm 3 finds an ε-optimal arm with probability at least $1-\delta$.*

Lemma 1.

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

and

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

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 2. *Suppose that $|\hat{\mu}_i^\ell - \mu_i^\ell| \leq \varepsilon/8$, $i = 0,1,2,3,4$. 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 $\mu_i^\ell \geq \mu_j^\ell$:

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

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

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

Proof.

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

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

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

□

Lemma 4. *Suppose that $|\hat{\mu}_i^\ell - \mu_i^\ell| \leq \varepsilon/8$, $i = 1,2,3$. If $\hat{\mu}_1^\ell, \hat{\mu}_3^\ell \in (\hat{\mu}_2^\ell - \varepsilon/4, \hat{\mu}_2^\ell + \varepsilon/4)$, 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^ℓ is closer to x^* than to x_2^ℓ, $t \leq 1/2$. Because of the concavity:

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

It 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 $x_2^\ell = (1-t) \cdot x^* + t \cdot x_3^\ell$. As x_2^ℓ is closer to x^* than to x_3^ℓ, $t \leq 1/2$. Because of the concavity:

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

It is 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^ℓ is closer to x^* than to x_1^ℓ, $1 > t \geq 1/2$. Because of the concavity:

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

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

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

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

□

By Lemma 1,

$$\mathbb{P}(\exists \ell,\ \exists i \in \{0,1,2,3,4\} : |\hat{\mu}_i^\ell - \mu_i^\ell| \geq \varepsilon/8) \leq \delta.$$

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 2 the optimal arm is never thrown away and by lemma 4 an ε-optimal arm is returned when the algorithm terminates. This way Algorithm 3 returns an ε-optimal arm with probability at least $1 - \delta$.

VI. Conclusion

I have investigated a special case of the multi-armed bandit problem, in which the arms are the elements of the $[0,1]$ interval and a concave function describes the expectations of the arms. I have developed an algorithm, that finds an ε-optimal arm with probability at least $1 - \delta$ in this case.

There are still many interesting problems that require further investigation including the case of infinitely many arms with a concave structure in higher-dimensional spaces.

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[2] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

[3] E. Even-Dar, S. Mannor, and Y. Mansour, "Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems," *Journal of Machine Learning Research*, vol. 7, no. 39, pp. 1079–1105, 2006.

[4] S. Mannor and J. N. Tsitsiklis, "The sample complexity of exploration in the multi-armed bandit problem," *Journal of Machine Learning Research*, vol. 5, no. Jun, pp. 623–648, 2004.

[5] B. C. Csáji and E. 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*. IEEE, 2011, pp. 3634–3639.