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 a new algorithm that finds the optimal arm with high probability and proved a theorem about its sample complexity.

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 this section we will consider a new problem that has not yet been investigated. 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.

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 $|S_\ell| > 1$ 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*

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

Lemma 1. *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:

$$\mathbb{P}\left(a_{i^*} \notin S_{\ell+1}\right) \leq \delta_\ell.$$

The $a_{i^*} \notin S_{\ell+1}$ if and only if $|i^* - i_\ell^*| > 2^{m-\ell-1}$, so:

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

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 $\mathbb{P}\left(i_\ell^* = i \cdot 2^{m-\ell-1}\right)$ 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 $i^* > i \cdot 2^{m-\ell-1}$.
Let $j = (i+1) \cdot 2^{m-\ell-1}$ and $i' = i \cdot 2^{m-\ell-1}$.
With these notations: $i^* > j > i' \implies \mu_{i^*} > \mu_j > \mu_{i'}$, because i^* is the index of the optimal arm in S_ℓ 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

$$\mu_j \geq \mu_{i'} + 2^{m-\ell-1}\Delta.$$

From the definition of i_ℓ^* follows that if $i_\ell^* = i'$ then $\hat{\mu}_{i'}^\ell \geq \hat{\mu}_j^\ell$.
This way: $\mathbb{P}(i_\ell^* = i') \leq \mathbb{P}(\hat{\mu}_{i'}^\ell \geq \hat{\mu}_j^\ell)$.
Consider the following events:

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

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

$$\begin{aligned}\hat{\mu}_j &> \mu_j - 2^{m-\ell-2}\Delta\\ &\geq \mu_{i'} + 2^{m-\ell-1}\Delta - 2^{m-\ell-2}\Delta\\ &= \mu_{i'} + 2^{m-\ell-2}\Delta\\ &> \hat{\mu}_{i'}\end{aligned}$$

This implies that:

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

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

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

Similarly,

$$\mathbb{P}(\overline{B}) \leq \delta_\ell/4.$$

The case when $i^* < i \cdot 2^{m-\ell-1}$ can be proved similarly. □

Lemma 2. *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

$$\mathbb{P}\left(i^* \neq i_m^*\right) \leq \delta_m.$$

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

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

$$\mathbb{P}(i_m^* = j) \leq \mathbb{P}(\hat{\mu}_j^m \geq \hat{\mu}_{i^*}^m)$$

By the definition of $i^* : \mu_j \le \mu_{i^*} - \Delta$.
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 $A \wedge B \implies C$:

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

This implies that:

$$\begin{aligned} \mathbb{P}(i^*_m = j) &\le \mathbb{P}(\hat{\mu}^m_j \ge \hat{\mu}^m_{i^*}) \\ &= \mathbb{P}(\overline{C}) \\ &\le \mathbb{P}(\overline{A} \vee \overline{B}) \\ &\le \mathbb{P}(\overline{A}) + \mathbb{P}(\overline{B}). \end{aligned}$$

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

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

Similarly,

$$\mathbb{P}(\overline{B}) \le \delta_m/4.$$

□

Lemma 3. *The sample complexity of Algorithm 2 is*

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

Proof. The number of arm samples in the ℓ-th round is $3\,n_\ell$.

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

□

Remark 2. *In the general case when $2^{m-1}+1 < n \le 2^m+1$ 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 $\hat{\mu}^1_j$ denote their empirical values and let $i^*_1 = \arg\max_j \hat{\mu}^1_j$. 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. Conclusion

A structured multi-armed bandit problem has been analyzed, in which 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

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

There are still many interesting problems that require further investigation including the case of infinitely many arms with a similar structure and the two-dimensional case when the arms are located on a grid.

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.