

Best Arm Identification in Bandits with Limited Precision Sampling

Ameya Bagal Aditya Mallick

Indian Institute of Technology, Madras

November 12, 2024

1 Introduction

- Problem description
- Applications

2 Problem Setup

3 Algorithms

- Boxed Bandit Modified Track and Stop (BBMTS)
- Boxed Bandit SEA (BBSEA)

4 Conclusion and Improvements

Limited precision sampling

- In contrast to classical multi-armed bandit problems, where we have direct control over arm pulls, this problem presents a scenario where our access to the arms is indirect.
- Here the learner samples the arms through exploration bundles (or 'boxes').
- At each time instant the learner chooses a box, and the boxes pulls an arm according to a box-specific probability distribution, which is unknown to the learner.
- The arm pulled and the instantaneous rewards are then revealed to the learner.

Applications and Related Setups

- This scenario can be interpreted as the case where we have limited control over arms because of external constraints.
- We can also use this to model the case where arm selection is noisy ('trembling hand' model or arm erasures).
- This model can also be seen as a way for the learner to preserve privacy. By exploring through the non adaptive selection profiles of the boxes, the learner can hide its own preferences from observers.

Problem Setup

We have a stochastic Multi Armed Bandit problem with,

- K arms, with the reward distribution ν_k and mean reward μ_k for each arm $k \in [K]$
- M boxes, where \mathcal{A}_m denotes the set of arms accesible through the box m .
- $q_{m,k}$ denotes the probability of choosing arm $k \in \mathcal{A}_m \subseteq K$ given the box $m \in M$ is chosen
- The tuple $C = (q, \nu)$ defines the problem instance, and the best arm is given by $a^*(C)$.
- π denotes any generic BAI policy. The stopping time is denoted by τ_π .
- For a given confidence $\delta \in (0, 1)$, our goal is to identify the best arm with probabily of error less than or equal to δ

Algorithm 1 - Track & Stop based algorithm

This algorithm is designed for the general setting where each arm may be associated with multiple boxes. The following assumptions are made

- The reward distribution of each arm is Gaussian with 0 mean and unit variance.
- Each arm is accessible through all M boxes. The underlying instance is denoted by $q_0 = \{q_{m,k}^0\}$ and $\mu_0 = \{\mu_k\}$.

Asymptotic Lower Bound

Theorem (1)

$$\liminf_{\delta \downarrow 0} \inf_{\pi \in \Pi(\delta)} \frac{\mathbb{E}[\tau_\pi]}{\log(1/\delta)} \geq \frac{1}{T^*(\mathbf{q}_0, \boldsymbol{\mu}_0)}, \quad (1)$$

where $T^*(\mathbf{q}_0, \boldsymbol{\mu}_0)$ in (1) is given by

$$T^*(\mathbf{q}_0, \boldsymbol{\mu}_0) = \sup_{w \in \Sigma_M} \inf_{\lambda \in \text{ALT}(\boldsymbol{\mu}_0)} \sum_{m=1}^M \sum_{k=1}^K w_m q_{m,k}^0 \frac{(\mu_k^0 - \lambda_k)^2}{2}, \quad (2)$$

Non unique Optimal Allocations

It is observed that there are multiple allocations which achieve the supremum in (2), making the optimal allocation non unique. We use $\mathcal{W}^*(q_0, \mu_0)$ to denote the set of all such allocations. The following result is obtained,

Lemma 1

The mapping $(q, \mu) \rightarrow \mathcal{W}^*(q, \mu)$ is upperhemicontinuous and compact-valued. Furthermore, $\mathcal{W}^*(q, \mu)$ is convex for all (q, μ) .

Challenges with tracking

- One of the major features of BAI algorithms is the convergence of empirical frequency of arm pulls to the optimal allocation.
- This becomes challenging in our case as instead of converging to one single optimal allocation, it might oscillate between multiple of the possible allocations.
- The paper proposes a strategy called the *modified D-tracking rule*, which demonstrates tracking like behaviour.

Modified D-tracking rule

Lemma 2

Let $f(t) = \frac{\sqrt{t}}{\sqrt{M}}$. Let $\{w(t)\}_{t=1}^{\infty} \subset \Sigma_M$ be any sequence such that $w(t+1) \in \mathcal{W}^*(\hat{\mathbf{q}}(t), \hat{\boldsymbol{\mu}}(t))$ for all t . Let $i_0 = 0$ and

$$i_{t+1} = (i_t \bmod M) + \mathbf{1}_{\{\min_{m \in [M]} N(t, m) < f(t)\}}, \quad t \geq 0.$$

Then, under the *modified D-tracking rule* given by

$$B_{t+1} = \begin{cases} i_t, & \min_{m \in [M]} N(t, m) < f(t), \\ b_t, & \text{otherwise,} \end{cases} \quad (3)$$

where $b_t = \arg \min_{m \in \text{supp}(\sum_{s=1}^t w(s))} N(t, m) - \sum_{s=1}^t w_m(s)$, we have

$$\lim_{t \rightarrow \infty} d_{\infty}((N(t, m)/t)_{m \in [M]}, \mathcal{W}^*(\mathbf{q}_0, \boldsymbol{\mu}_0)) = 0 \quad \text{a.s..} \quad (4)$$

Modified D-tracking rule

- Under this rule, it is ensured that the boxes are sampled in a round-robin fashion till, $N(t, m) = \Omega(t)$ where $N(t, m)$ is the number of times box m has been sampled till time t .
- After that, the boxes are sampled based on the previous allocations $w(s) : 1 \leq s \leq t$.
- It is proved that the empirical average of all the allocations up to time t , $\bar{w}(t) = 1/t \sum_{s=1}^t w(s)$, approaches $\mathcal{W}^*(\mathbf{q}_0, \mu_0)$ as $t \rightarrow \infty$.
- Similarly, because of the upperhemicontinuity of \mathcal{W}^* , $\mathcal{W}^*(q(t), \mu(t))$ is close to $\mathcal{W}^*(q_0, \mu_0)$ for a large t .

Generalised Likelihood Ratio Test

Lemma 3

Let $Z_{a,b}(t)$ denote the generalised likelihood ratio test statistic. Fix $a, b \in [K]$ and a policy π . Fix t such that $\min_{k \in [K]} N_k(t) > 0$ a.s.. If $\hat{\mu}_a(t) \geq \hat{\mu}_b(t)$, then

$$Z_{a,b}^\pi(t) = N_a(t) \frac{(\hat{\mu}_a(t) - \hat{\mu}_{a,b}(t))^2}{2} + N_b(t) \frac{(\hat{\mu}_b(t) - \hat{\mu}_{a,b}(t))^2}{2}, \quad (5)$$

where $\hat{\mu}_{a,b}(t)$ is defined as

$$\hat{\mu}_{a,b}(t) := \frac{N_a(t)}{N_a(t) + N_b(t)} \hat{\mu}_a(t) + \frac{N_b(t)}{N_a(t) + N_b(t)} \hat{\mu}_b(t). \quad (6)$$

Algorithm: Boxed-Bandit Modified Track-and-Stop

Input

$\delta \in (0, 1)$, $\rho > 0$, $K \in \mathbb{N}$, and $M \in \mathbb{N}$

Output

$\hat{a} \in [K]$ – best arm

Initialisation: $t = 0$, $\hat{\mu}_a(t) = 0$ for all $a \in [K]$, $Z(0) = 0$

1. Compute $\hat{a}(t) = \arg \max_a \hat{\mu}_a(t)$.
2. **If** $Z(t) \geq \zeta(t, \delta, \rho)$ and $\min_k N_k(t) > 0$:
 - Stop box selections.
 - **Return** $\hat{a} = \arg \max_{a \in [K]} \hat{\mu}_a(t)$.
3. **Else:**
 - Select box B_{t+1} as per *modified D-tracking rule*.
 - Update $\hat{\mathbf{q}}(t)$, $\hat{\boldsymbol{\mu}}(t)$, and $Z(t)$.
 - Go to step 1.

Stopping Condition

The stopping condition for the algorithm is defined as follows:

- Stop if

$$Z(t) := \min_{b \neq \hat{a}(t)} Z_{\hat{a}(t), b}(t) \geq \zeta(t, \delta, \rho) \quad (7)$$

- where the threshold $\zeta(t, \delta, \rho)$ is given by

$$\zeta(t, \delta, \rho) := \log \left(\frac{C t^{1+\rho}}{\delta} \right)$$

where C is a constant that satisfies

$$\sum_{t=1}^{\infty} \frac{e^{K+1}}{K^K} \frac{(\log^2(C t^{1+\rho}) \log t)^K}{t^{1+\rho}} \leq C.$$

Performance analysis of BBMTS algorithm

- ① $\pi_{\text{BBMTS}}(\delta, \rho) \in \Pi(\delta)$ for each $\delta \in (0, 1)$ and $\rho > 0$.
- ② For each $\rho > 0$, the stopping time of $\pi_{\text{BBMTS}}(\delta, \rho)$ satisfies

$$\limsup_{\delta \downarrow 0} \frac{\tau_{\pi_{\text{BBMTS}}(\delta, \rho)}}{\log(1/\delta)} \leq \frac{1 + \rho}{T^*(\mathbf{q}_0, \boldsymbol{\mu}_0)} \quad \text{a.s..} \quad (8)$$

- ③ For each $\rho > 0$, the quantity $\mathbb{E}[\tau_{\pi_{\text{BBMTS}}(\delta, \rho)}]$ satisfies

$$\limsup_{\delta \downarrow 0} \frac{\mathbb{E}[\tau_{\pi_{\text{BBMTS}}(\delta, \rho)}]}{\log(1/\delta)} \leq \frac{1 + \rho}{T^*(\mathbf{q}_0, \boldsymbol{\mu}_0)}. \quad (9)$$

Partitioned Arms

- Consider the case where each arm is only present in a single box. In other words, the arms are partitioned into the boxes.
- The set of arms \mathcal{A}_m that are accessible using box m is known, but not the arm selection probabilities $(q_{m,k})$ within the boxes.

Ex. $K = 9, M = 3$

Box 1

$$A_0 = (a_{0,0}, a_{0,1}, a_{0,2})$$

$$q_0 = (0.3, 0.4, 0.4)$$

Box 2

$$A_1 = (a_{1,0}, a_{1,1}, a_{1,2})$$

$$q_1 = (0.2, 0.3, 0.6)$$

Box 3

$$A_2 = (a_{2,0}, a_{2,1}, a_{2,2})$$

$$q_2 = (0.1, 0.4, 0.5)$$

Boxed-Bandit Successive Elimination Algorithm

Input: $K, M, \delta > 0, \mathcal{A}_m$ for $m \in [M]$

Output: $\hat{a}_{\text{BBSEA}} \in [K]$ (best arm)

Initialisation: $n, t = 0, S_m = \mathcal{A}_m \forall m, S = \bigcup_m S_m, \hat{\mu}_{m,k}(0) = 0 \forall m, k.$

Algorithm:

- While $|S| > 1$
 - ① $n \leftarrow n + 1$
 - ② Select each box until every active arm $A_{m,k}$ in box m is pulled **at least** n times. For every box selection, increment t by 1.
 - ③ Update empirical mean, UCB and LCB for all active arms.
 - ④ For every box m , eliminate all arms in S_m having UCB lower than LCB of at least one other arm in S .
 - ⑤ $S \leftarrow \bigcup_m S_m$
- Output \hat{a}_{BBSEA}

Working of BBSEA

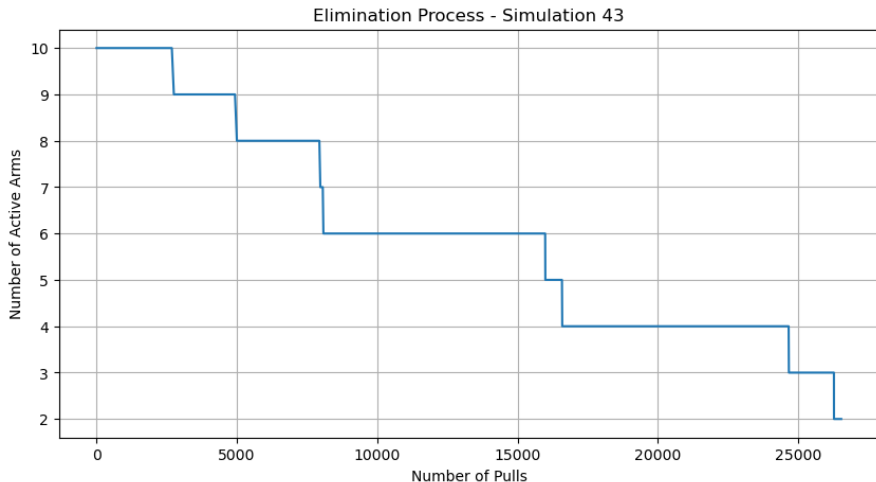
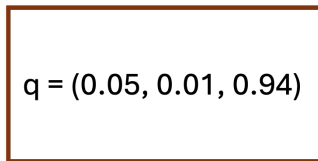


Figure: Algorithm Visualisation ($M = 3$, $K = 10$)

Analysis of BBSEA

- The main difference from SEA arises from the unknown arm probabilities within boxes.
- We cannot ensure that we pick each arm a fixed number of times every round, since we do not know which arm we will pick from a given box.
- Try to pull each arm a minimum number of times within a round.
- May become a problem in the case of skewed boxed-bandit distributions.



$q = (0.05, 0.01, 0.94)$

Figure: Box with skewed arms

Theoretical Results

① Define:

$$\alpha_{m,k} = 1 + \frac{102}{\Delta_{m,k}^2} \log \left(\frac{64 \sqrt{\frac{8K}{\delta}}}{\Delta_{m,k}^2} \right) \quad (10)$$

for arm k , in box m .

② Moreover, let

$$\beta_{m,k} = \frac{1}{q_{m,k}^0} \left[\alpha_{m,k} + 2 \log \frac{2K}{\delta} + 2 \sqrt{\log \frac{2K}{\delta} \left(\log \frac{2K}{\delta} + \alpha_{m,k} \right)} \right] \quad (11)$$

and let $\beta_m = \max_{k \in \mathcal{A}_m} \beta_{m,k}$.

Theoretical Results (contd.)

Theorem

For 1-subgaussian rewards, and a fixed confidence $\delta \in (0, 1)$, the following hold with probability greater than $1 - \delta$:

- ① *BBSEA outputs the best arm correctly.*
- ② *The stopping time of BBSEA is $\leq \sum_{m=1}^M \beta_m$.*
- ③ *Furthermore, for any policy π , a lower bound on the stopping time is given by:*

$$\mathbb{E}[\tau_\pi] \geq \log\left(\frac{1}{2.4\delta}\right) \cdot \sum_{m=1}^M \max_{k \in \mathcal{A}_m} \frac{1}{q_{m,k}^0 \Delta_{m,k}^2}. \quad (12)$$

Simulations

```
# Define two q variations
q_variations = {
    'Uniform': {
        0: [0.33, 0.33, 0.34],
        1: [0.25, 0.25, 0.25, 0.25],
        2: [0.33, 0.33, 0.34]
    },
    'Skewed': {
        0: [0.01, 0.02, 0.97],
        1: [0.02, 0.03, 0.87, 0.08],
        2: [0.98, 0.01, 0.01]
    }
}
```

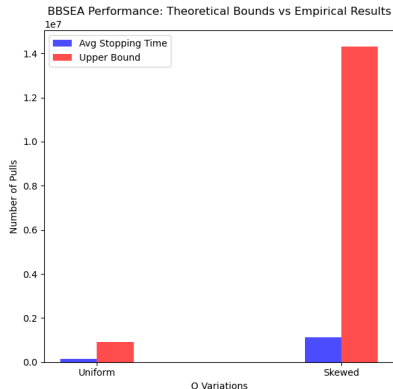


Figure: Skewed boxed bandits

Simulations (contd.)

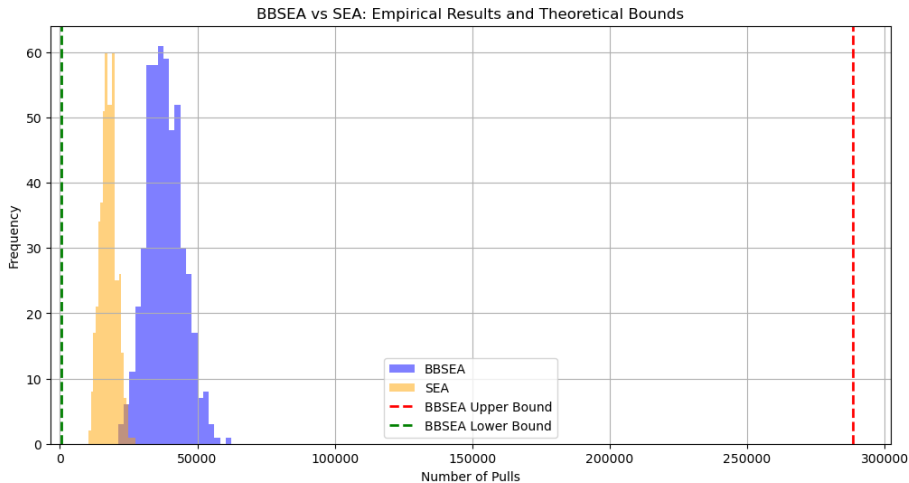


Figure: 500 runs ($M = 3$, $K = 10$)

Simulations (contd.)

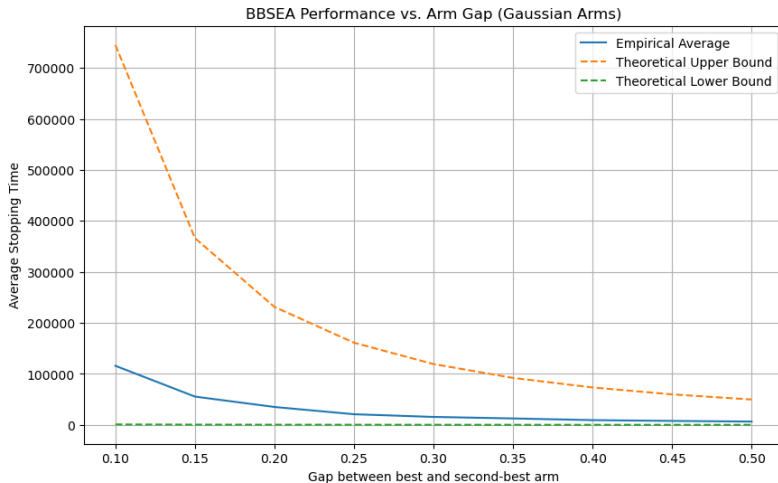


Figure: $M = 3$, $K = 10$

Simulations (contd.)

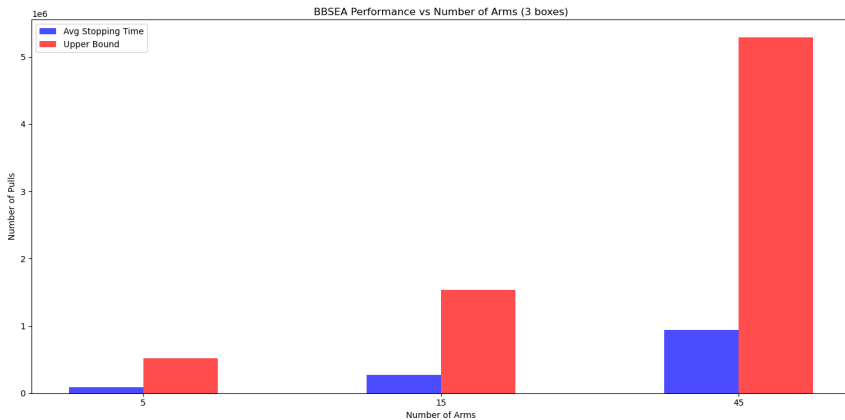


Figure: $M = 3$

Conclusion and Improvements

- The authors showed that asymptotic optimality can be achieved by tracking the empirical average of arbitrarily chosen optimal allocations.
- The paper proposes an algorithm achieving a non-asymptotic bound for the special case of partitioned arms.
- We elaborate on a small drawback of this algorithm when dealing with the skewed-bandit problem.
- Instead of SEA for partitioned arms, we can consider using better BAI algorithms such as lil'UCB and Exponential-Gap Elimination and adapting them to the problem setting to achieve a tighter upper bound.

Thank You!