Asymptotically Optimal Fixed-Budget Best Arm Identification with Variance-Dependent Bounds

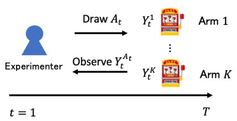
Masahiro Kato, The University of Tokyo
Masaaki Imaizumi, The University of Tokyo
Takuya Ishihara, Tohoku University
Toru Kitagawa, Brown University

1. ABSTRACT

➢ Best arm identification (BAI) with a fixed budget.
- Recommend the best arm at the end of an experiment.
- Minimize the expected simple regret.
  Regret computed only at the end of an experiment.
➢ **Contribution: Minimax optimal strategy for BAI.**
- Capture the uncertainty by the worst-case analysis.
- Derive variance-dependent tight lower and upper bounds.

2. PROBLEM SETTING

- **Adaptive experiment with T rounds**: $[T] = \{1,2,\dots,T\}$.
- **K treatment arms**: $[K] = \{1,2,\dots,K\}$.
  Treatment arms: alternatives of medicine, policy, and advertisements.
- Each arm a has a potential outcome $Y_t^a \in \mathbb{R}$.
- The distributions of Y_t^a do not change.
- $P = (P^1,..,P^K)$: distribution of $(Y_t^1,\dots,Y_t^K)$. $\mathcal{P}$: a set of P.
- Denote the expectation and probability laws under P by $\mathbb{E}_P$ and $\mathbb{P}_P$.
- Denote the mean outcome of an arm a by $\mu^a(P) = \mathbb{E}_P[Y_t^a]$.
- Best treatment arm: an arm with the highest reward.
- Denote the best treatment arm by $a^*(P) = \arg\max_{a\in[K]} \mu^a(P)$
- **Sequence of an experiment**: at round t
- Draw a treatment arm $A_t \in [K]$.
- Observe a reward $Y_t^{A_t}$
- After the final round T, an algorithm recommends an estimated best treatment arm $\hat{a}_T \in [K]$.



- **Goal**: Minimize the expected simple regret.

3. EVALUATION

- Simple regret: $r_T(P) = \mu^{a^*(P)}(P) - \mu^{\hat{a}_T}(P)$ under P
➢ **Expected simple regret**: $\mathbb{E}_P[r_t(P)] = \mathbb{E}_v\left[\mu^{a^*(P)}(P) - \mu^{\hat{a}_T}(P)\right]$.
➢ Our goal is to minimize the expected simple regret.

4. WORST-CASE LOWER BOUNDS

- To capture the uncertainty of the problem, we consider the worst-case expected simple regret.
- $\Delta^{a,b}(P) = \mu^a(P) - \mu^b(P)$.
- For some constant $C > 0$, the expected regret becomes

$$\mathbb{E}_P[r_t(P)] = \sum_{b\neq a^*(P)} \Delta^{a^*(P),b}(P) \exp\left(-CT\left(\Delta^{a^*(P),b}(P)\right)^2\right).$$

- The worst-case gap for P is $\Delta^{a^*(P),b}(P) = \frac{C_1}{\sqrt{T}}$, for some C_1.

→ For some constant $C_2 > 0$, $\max_{P\in\mathcal{P}} \mathbb{E}_P[r_t(P)] \approx \frac{C_2}{\sqrt{T}}$.

➢ Lower bounds.
- Tight lower bound characterized by the variances of outcomes.
- $(\sigma^a)^2$: Variance of Y_t^a.
- Assume that for all $P \in \mathcal{P}$, the variances are the same.
- For any $K \geq 2$, any strategy (algorithm) satisfies

$$\max_{P\in\mathcal{P}} \mathbb{E}_P[r_t(P)] \geq \frac{1}{12}\sqrt{\sum_{a\in[K]} (\sigma^a)^2}. \quad (1)$$

- We can refine the lower bounds for a case with $K = 2$.
- If $K = 2$, any strategy satisfies

$$\max_{P\in\mathcal{P}} \mathbb{E}_P[r_t(P)] \geq \frac{1}{12}\sqrt{(\sigma^1 + \sigma^2)^2}. \quad (2)$$

- Note that $\sum_{a\in[K]}(\sigma^a)^2 \geq (\sigma^1 + \sigma^2)^2$.

➢ **Target allocation ratio**.
- A ratio of $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T A_t\right]$ that a strategy want to achieve.
- Define $w^*: [K] \to [0,1]$ such that $\sum_{a\in[K]} w^*(a) = 1$.
- From the lower bounds, we can derive a target allocation ratio.
- To achieve the lower bound in (1), $w^*(a) = \frac{(\sigma^a)^2}{\sum_{b\in[K]}(\sigma^b)^2}$.
- To achieve the lower bound in (2), $w^*(a) = \frac{\sigma^a}{(\sigma^1+\sigma^2)^2}$.

5. TS-SA STRATEGY

- **TS**: two-stage sampling.
- **SA**: recommendation using the sample average estimator.

➢ **Procedure of the TS-SA strategy.**
- **Split T rounds into two stages.**
- **At the first stage**, we uniformly randomly draw arms.
- Estimate w^* by replacing $(\sigma^a)^2$ with a sample variance $(\hat{\sigma}^a)^2$.
  When $K = 2$, we use $\hat{w}(a) = \frac{\hat{\sigma}^a}{(\hat{\sigma}^1+\hat{\sigma}^2)^2}$.
  When $K \geq 3$, we use $\hat{w}(a) = \frac{(\hat{\sigma}^a)^2}{\sum_{b\in[K]}(\hat{\sigma}^b)^2}$.
- **At the second stage,** we draw (A_t) following an estimate $\hat{w}$ of w^*.
- At the end of the process, recommend $\hat{a}_T = \arg\max_{a\in[K]} \hat{\mu}_T^a$.
- $\hat{\mu}_T^a$ is the sample average defined as

$$\hat{\mu}_T^a = \frac{1}{\sum_{t=1}^T 1[A_t = a]} \sum_{t=1}^T 1[A_t = a] Y_t^a.$$

6. WORST-CASE UPPER BOUNDS

- Worst-case regret upper bound of the TS-SA strategy.
- When $K = 2$, the TS-SA strategy satisfies

$$\max_{P\in\mathcal{P}} \mathbb{E}_P[r_t(P)] \leq \frac{1}{2}\sqrt{(\sigma^1 + \sigma^2)^2}.$$

- When $K \geq 3$, the TS-SA strategy satisfies

$$\max_{P\in\mathcal{P}} \mathbb{E}_P[r_t(P)] \leq 2\log K \sqrt{\sum_{a\in[K]} (\sigma^a)^2}.$$

The leading factor of the upper bounds matches the lower bounds.

7. Contextual Information

- Utilize contextual information X_t observed before drawing an arm.
- We can enhance the efficiency of strategies.

References

Bubeck, S., Munos, R., and Stoltz, G. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 2011

Kaufmann, E., Cappe, O., and Garivier, A. (2016), "On the Complexity of Best-Arm Identification in Multi-Armed " Bandit Models," JMLR, 17, 1-42