Adversarial Best Arm Identification in Gaussian Bandits

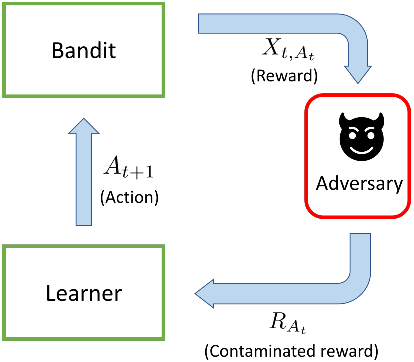
Arpan Mukherjee

Department of Electrical, Computer, and Systems Engineering

Rensselaer Polytechnic Institute

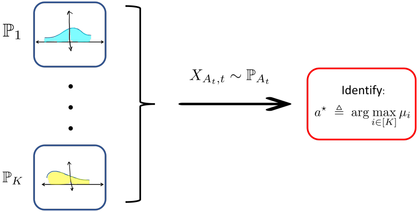
April 19, 2022

Introduction



- *Contaminated* best arm identification (CBAI)
- Adversary may contaminate reward sample
- Identify the arm with largest *mean*

Best Arm Identification (BAI)



- K-armed Gaussian bandit $\{\mathcal{N}(\mu_i, \sigma^2) : i \in [K]\}$
- Each arm has unknown mean μ_i
- **Objective:** Identify arm with largest mean

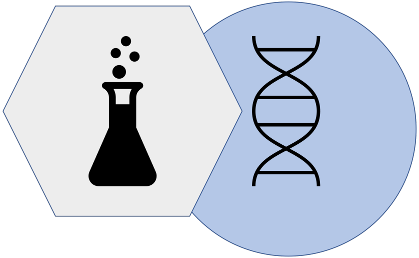
$$a^\star \triangleq \arg\max_{i\in[K]} \mu_i$$

- **Fixed confidence setting:**
  - Minimize the sample complexity
  - Constraint on the probability of error

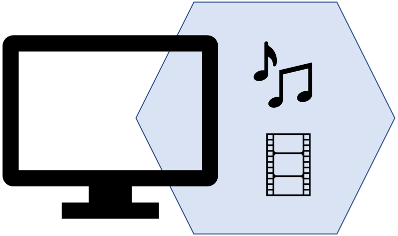
$$\mathbb{P}(\hat{a} \neq a^\star) < \delta$$

Rensselaer

BAI: Applications

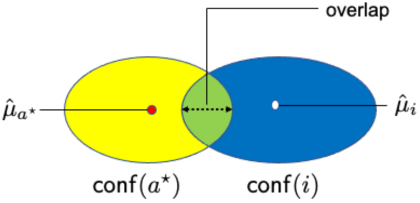


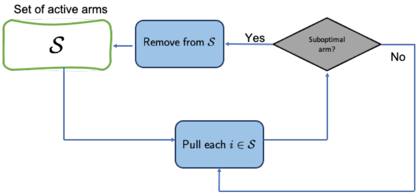
- Drug identification in clinical trials



- Recommendation system

BAI: Existing Algorithms

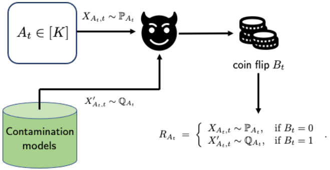




- Confidence interval based approach
- Construct confidence interval around mean estimates
- Continue sampling till overlap vanishes
- Representative studies:
  - Gabillon, NeurIPS'12
  - Jamieson, COLT'14
  - Kalyankrishnan, ICML'12
  - Kaufmann, JMLR'13

- Successive elimination based approach
- Maintain an active set of arms
- Remove suboptimal arms
- Continue till one arm remains
- Representative studies:
  - Audibert, COLT'10
  - Chen, COLT'17
  - Even-Dar, JMLR'06

Contaminated Best Arm Identification (CBAI)



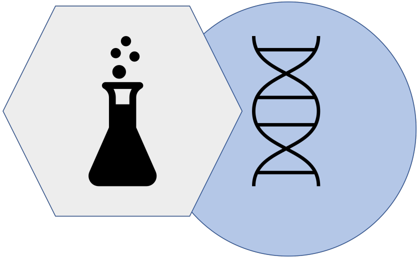
- At every arm pull, the adversary flips a coin $B_t \sim \text{Bern}(\varepsilon)$
- If the outcome is "tails"(0), then the true reward $X_{A_t,t} \sim \mathbb{P}_{A_t}$ is sent to the learner
- Otherwise, a random sample from a contamination model $X'_{A_t,t} \sim \mathbb{Q}_{A_t}$ is sent to the learner
- **Relaxed constraint on decision error:**

$$\boxed{\mathbb{P}(\mu_{\hat{a}} < \mu_{a^\star} - U) < \delta}$$

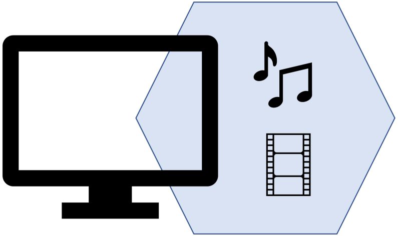
Definition (Oblivious adversary)

The adversary is defined as oblivious if we have that for all $i \in [K]$, the tripples $\{X_{i,t}, X'_{i,t}, B_t\}_{t\geq 1}$ are assumed to be independent of each other.

Motivation



- Efficacy of drug response in clinical trials
- Fraction of results reported incorrectly
- Fraction of samples contaminated

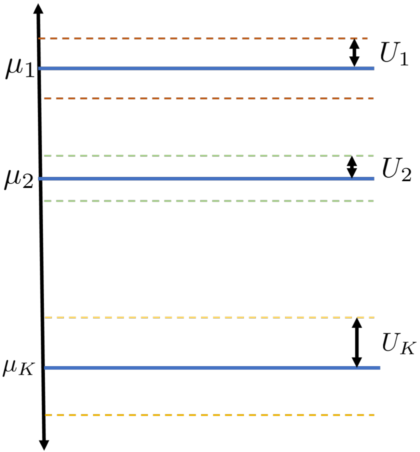


- Recommendation system
- Recommendation based on user feedback
- User feedback could be malicious or spam

Related Work

- Contaminated stochastic bandits for regret minimization [Lykouris, SIGACT'19]
  - adversarial power characterized by total number of corrupted samples C
  - algorithm proposed by this study degrades linearly with C, which is the optimal rate
- For best arm identification(BAI), the problem was first studied in [Altschuler, JMLR'19]
  - Best arm redefined as the arm with largest *median*
  - Three adversarial models: *Oblivious* adversary, *prescient* adversary and *malicious* adversary
  - Considers very restrictive class of cumulative distribution functions (CDFs)
  - Results not valid for all discrete models (like Bernoulli bandits) or heavy-tailed continuous models
- Mean-based contaminated BAI was first investigated in [Mukherjee, NeurIPS'21]
  - Gap-based and successive elimination based algorithms proposed
  - Asymptotically optimal up to constant factors in sub-Gaussian bandits

Partially Identifiable Best Arm Identification (PIBAI)



- Arm means cannot be estimated *exactly*
- Estimation up to uncertainty U_i around mean
- CBAI is a special case of PIBAI
  - U_i's depend on σ and ε

- **Assumption 1**: no overlapping intervals:

$$(\mu_{a^\star} - U_{a^\star}) > (\mu_a + U_a)\ , a \in [K] \setminus a^\star$$

- **Assumption 2**: Corruption level $\varepsilon < 1/2$

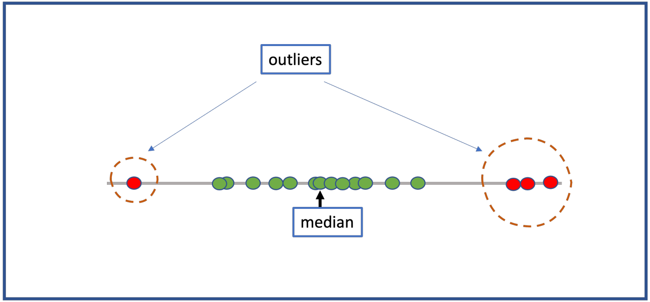
robust Top Two Sequential Probability Ratio Test (rTT-SPRT):

Overview:

- Likelihood ratio (LLR)-based approach
- Score-based outlier removal
- Randomized sampling between the top two arms
- Stop collecting samples when sufficient evidence has been collected to distinguish the top two arms

Scoring mechanism

- Rewards from arm $i \in [K]$: $Y_i^t \triangleq \{R_s, s \in [t] : A_s = i\}$
- Empirical median of arm $i \in [K]$ at time t: $\mathrm{med}(Y_i^t)$
- Score for each sample: $Z_{A_t,t} \triangleq \frac{|R_{A_t} - \mathrm{med}(Y_{A_t}^t)|}{\sigma}$



Rewards from arm $i \in [K]$

rTT-SPRT Algorithm: Notations

- For each arm $i \in [K]$, compute filtered rewards: $\tilde{Y}_i^t$ by removing the ϵ fraction of largest scores

- For any $(i,j) \in [K] \times [K]$, compute generalized log-likelihood ratio (GLLR):

$$\Lambda_t(i,j) \triangleq \log \frac{\max\limits_{\mu_i > \mu_j} f_i(\tilde{Y}_i^t \mid \mu_i) f_j(\tilde{Y}_j^t \mid \mu_j)}{\max\limits_{\mu_j > \mu_i} f_i(\tilde{Y}_i^t \mid \mu_i) f_j(\tilde{Y}_j^t \mid \mu_j)}$$

- Sample mean of the filtered sequence from $i \in [K]$: $\mu_{t,i} \triangleq \frac{1}{|\tilde{Y}_i^t|} \sum\limits_{y \in \tilde{Y}_y^t} y$

- Closed form for Gaussian bandits:

$$\Lambda_t(i,j) \triangleq \frac{(\mu_{t,i} - \mu_{t,j})^2}{2\sigma^2} \mathbb{1}_{\{\mu_{t,i} > \mu_{t,j}\}}$$

rTT-SPRT: Sampling and Stopping Rules

- **Arm selection rule.** Select randomly between the top two arms
  - Top arm: $I_1^t \triangleq \arg\max_{i \in [K]} \mu_{t,i}$
  - Second arm: $I_2^t \triangleq \arg\min_{i \in [K] \setminus \{I_1^t\}} \Lambda_t(I_1^t, i)$
  - Flip a coin $D_t \sim \text{Bern}(\beta)$,

$$\boxed{A_{t+1} \triangleq \begin{cases} I_1^t\,, & \text{if } \; D_t = 1 \\ I_2^t\,, & \text{if } \; D_t = 0 \end{cases}}$$

- **Stopping rule.** Stop as soon as top-two arms are sufficiently distinguishable

$$\boxed{\tau \triangleq \inf \left\{ t \in \mathbb{N} : \Lambda_t(I_1^t, I_2^t) > c_{t,\delta} \right\} .}$$

δ-PAC Guarantee

Theorem

For any $\delta \in (0,1)$, rTT-SPRT is δ-PAC in the PIBAI setting for the choice of the threshold

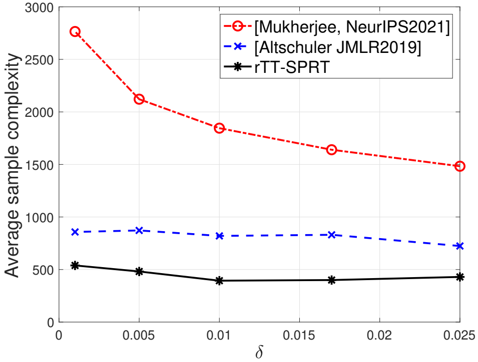
$$c_{t,\delta} \triangleq \frac{1}{(1-\varepsilon)^2} \log \frac{(K-1)Ct^{\alpha}}{\delta},$$

for every $a \in [K]$ and any $\alpha > 1$, where C is a constant set as $C \triangleq (1 + (\alpha-1)^{-1})$.

- Non-contaminated setting suffices with $c_{t,\delta} \triangleq \log \frac{(K-1)Ct^{\alpha}}{\delta}$
- $\mathcal{O}(1/(1-\varepsilon)^2)$ factor for dealing with contamination

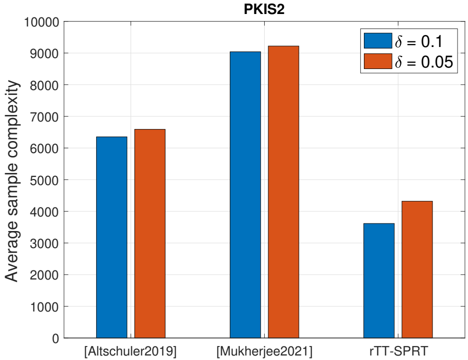
Experimental Result: Synthetic Data

- Model: $\boldsymbol{\mu} \triangleq [5, 4.5, 1, 1, 1]$
- Adversary: $\varepsilon = 0.1$
- $c_{t,\delta} \triangleq \frac{1}{(1-\varepsilon)^2} \log \frac{2(K-1)t^2}{\delta}$
- Tuning parameter $\beta = 0.5$
- $\sigma = 1$
- Averaged over 200 independent trials



Experimental Result: Real Data

- PKIS2: protein kinase + kinase inhibitors
- **Goal**: find kinase inhibitors for treating cancer cells
- We test $K = 4$ inhibitors against "ACVRL1" kinase
- Logarithm of percentage control assumes a Gaussian distribution
- Averaged over 200 independent trials
- rTT-SPRT requires significantly fewer samples for indentifying the best inhibitor



Conclusions

- BAI in contaminated Gaussian bandits
- Score-based outlier removal, SPRT based arm selection
- Stopping rule meets the δ-PAC guarantee
- Empirically superior performance on synthetic and real world data

Future Directions

- Sample complexity analysis of rTT-SPRT
- Extension to general bandits
- Dealing with more powerful adversaries (prescient + malicious)
- Time uniform concentration results for tighter confidence intervals