Best Arm Identification for Bandits with Shifting Means

Lukas Zierahn, Wouter M. Koolen, Christina Katsimerou, Shubhada Agrawal and Dirk van der Hoeven

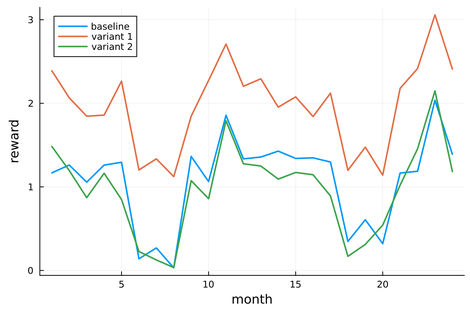
CWI Centrum Wiskunde & Informatica

UNIVERSITY OF TWENTE.

Summary

We study Best Arm Identification
We model non-stationarity as **adversarial mean shifts**
We develop active testing algorithms with **e-variables**
We prove optimal sample complexity

Real rewards are not stationary



Shifting Means Model

We model the reward of arm a in round t by

$$Y_{t,a} \;\coloneqq\; \mu_a + B_t + \epsilon_{t,a}$$

- μ_a is the base arm mean, fixed across time
- B_t is a time-varying shift, common to all arms
- $\epsilon_{t,a}$ is centred σ^2-sub-Gaussian noise

Implies **time-invariant** best arm $a^* \coloneqq \arg\max_a \mu_a$ and suboptimality gaps $\Delta_a \coloneqq \mu_{a^*} - \mu_a$.

Assume known **shifted mean bound** $|\mu_a + B_t| \le U$

Protocol

Bandit hides base means $\mu_1, \dots, \mu_K$.
for $t = 1, 2, \dots, \tau$ **do**
  Learner samples arm $A_t \sim w_t$
  Bandit samples shift B_t and noises $\epsilon_{t,1} \cdots \epsilon_{t,K}$
  Learner observes only reward $X_t \coloneqq Y_{t,A_t}$
end for
Learner recommends arm $\tilde{a}$.

Data stream $(A_1, X_1), (A_2, X_2), \dots, , (A_\tau, X_\tau), \tilde{a}$

Design problem

A strategy for Learner is specified by sampling rule w_t, stopping rule τ, recommendation rule $\tilde{a}$.

It is δ-PAC if $\mathbb{P}_{\mathcal{B}}\{\tau < \infty \text{ and } \tilde{a} \neq a^*(\mathcal{B})\} \le \delta$ against every shifting bandit $\mathcal{B}$.

It is efficient if its sample complexity $\mathbb{E}[\tau]$ is small.

Novel e-values

Say the learner picks $A_t \sim w_t$. For any pair $a \neq b$ of arms, we base everything on the statistic

$$s_{t,a,b} \;\coloneqq\; e^{z_{t,a,b}} \quad \text{where} \quad z_{t,a,b} \;\coloneqq\; \frac{\mathbb{I}[A_t = a]}{w_{t,a}}\left(\lambda_{t,a,b} X_t - \alpha_{t,a,b}\right) + \frac{\mathbb{I}[A_t = b]}{w_{t,b}}\left(-\lambda_{t,a,b} X_t - \beta_{t,a,b}\right)$$

We choose parameters $(\lambda_{t,a,b}, \alpha_{t,a,b}, \beta_{t,a,b})$ to make

- $s_{t,a,b}$ an **e-variable** against $\{\mu_b \ge \mu_a\}$, i.e. $\mathbb{E}_t[s_{t,a,b}] \le 1$, and hence
- $S_{t,a,b} \coloneqq \prod_{i=1}^t s_{i,a,b}$ a **supermartingale** against $\{\mu_b \ge \mu_a\}$, and consequently
- $P_{t,a} \coloneqq \min_{b \neq a} S_{t,a,b}$ an **e-process** against $\bigcup_{b \neq a}\{\mu_b \ge \mu_a\} = \{a \neq a^*\}$.

We stop at time $\tau \;\coloneqq\; \inf\left\{ t \in \mathbb{N} \,\middle|\, \max_{a \in [K]} P_{t,a} \;\ge\; \frac{K-1}{\delta} \right\}$ and recommends arm $\tilde{a} \;\coloneqq\; \arg\max_{a \in [K]} P_{\tau,a}$.

Safety and E-power

$$\max_{\mu_b \ge \mu_a} \mathbb{E}_t[s_{t,a,b}] \;=:\; \Phi(w_{t,a}, w_{t,b}, \lambda_{t,a,b}, \alpha_{t,a,b}, \beta_{t,a,b})$$

$$\mathbb{E}_t[z_{t,a,b}] \;=\; \lambda_{t,a,b}(\mu_a - \mu_b) - \alpha_{t,a,b} - \beta_{t,a,b}$$

where the maximum expectation under the null is

$$\Phi(w_a, w_b, \lambda, \alpha, \beta) \;=\; \max_{u \in \{\pm U\}} w_a \exp\left(\frac{\lambda u - \alpha}{w_a} + \frac{\lambda^2 \sigma^2}{2 w_a^2}\right) + w_b \exp\left(\frac{-\lambda u - \beta}{w_b} + \frac{\lambda^2 \sigma^2}{2 w_b^2}\right) + 1 - w_a - w_b.$$

Sampling to Maximise E-power

Single one-round growth-rate optimisation problem to tune **everything**

$$D^*(\Delta) \;\coloneqq\; \max_{w \in \triangle_K} \min_{b \neq a^*} \max_{\substack{\lambda,\alpha,\beta \\ \Phi(w_{a^*}, w_b, \lambda, \alpha, \beta) \le 1}} \lambda \Delta_b - \alpha - \beta$$

Computationally efficient: **Jointly concave** in sampling rule w and all e-value parameters $(\lambda_b, \alpha_b, \beta_b)_{b \neq a^*}$.

Main result

Two standard techniques:

- We force $1/\sqrt{t}$ exploration in round t.
- We plug estimate for Δ into D^*

Our strategy for Learner is δ-PAC and satisfies

$$\mathbb{E}[\tau] \;\le\; \frac{\ln \frac{1}{\delta}}{D^*(\Delta)} + o(\ln \frac{1}{\delta}).$$

Reflections

Interpretation of our upper bound:

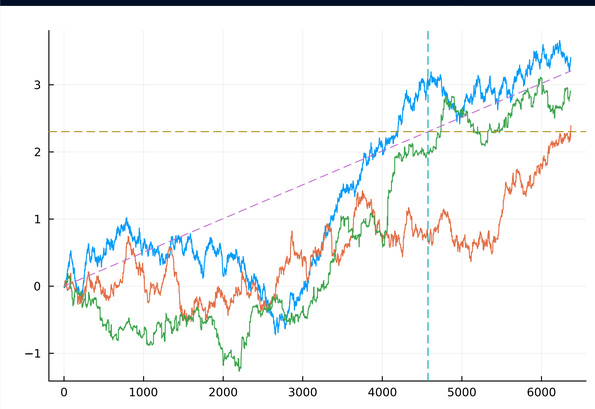
$$\frac{1}{D^*(\Delta)} \;\propto\; \sum_{a \neq a^*} \frac{\sigma^2 + U^2}{\Delta_a^2}$$

Nearly-matching lower bound: any δ-PAC algorithm must have

$$\lim_{\delta \to 0} \frac{\mathbb{E}[\tau]}{\ln \frac{1}{\delta}} \;\ge\; \Omega\left(\sum_{a \neq a^*} \frac{\sigma^2 + U^2}{\Delta_a^2}\right).$$

We show TAS and GLRT fail (due to (a) any deterministic fails and (b) overfitting to shifts)

Proofs rely on anytime-valid Bernstein.

Open: what if gaps are changing? Safety ✓Power ?

What $\ln S_{t,a^*,b}$ look like