CWI

Best Arm Identification for Shifting Means with Uniform Sampling

Lukas Zierahn[1, 2] Wouter M. Koolen[1, 3] Dirk van der Hoeven [4]
Shubhada Agrawal[5] Christina Katsimerou[2]

CWI[1] Booking.com[2] Twente University[3] Leiden University[4] Indian Institute of Science[5]

Booking.com

Best Arm Identification - Basics

Best Arm Identification (BAI) aims to find the best action from a set of actions $\mathcal{A}$ of size K interacting with the environment as follows,

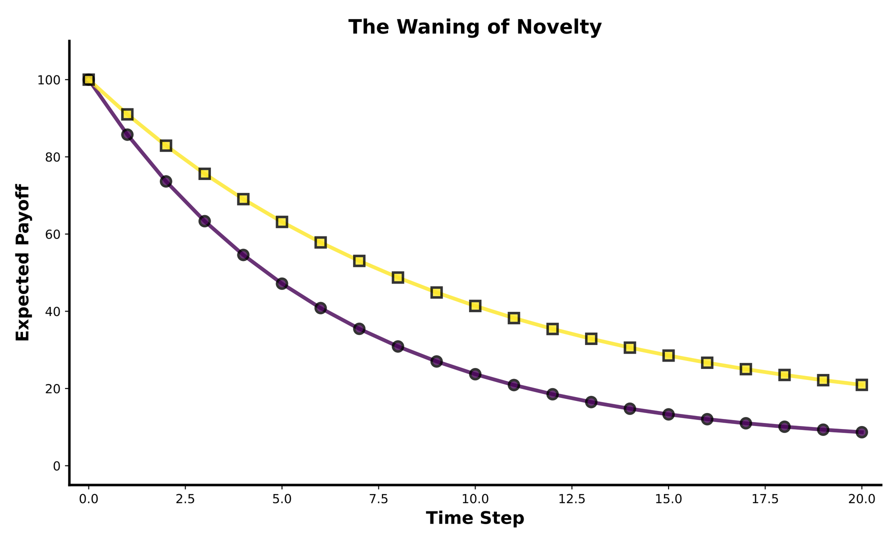
Vanilla Best Arm Identification

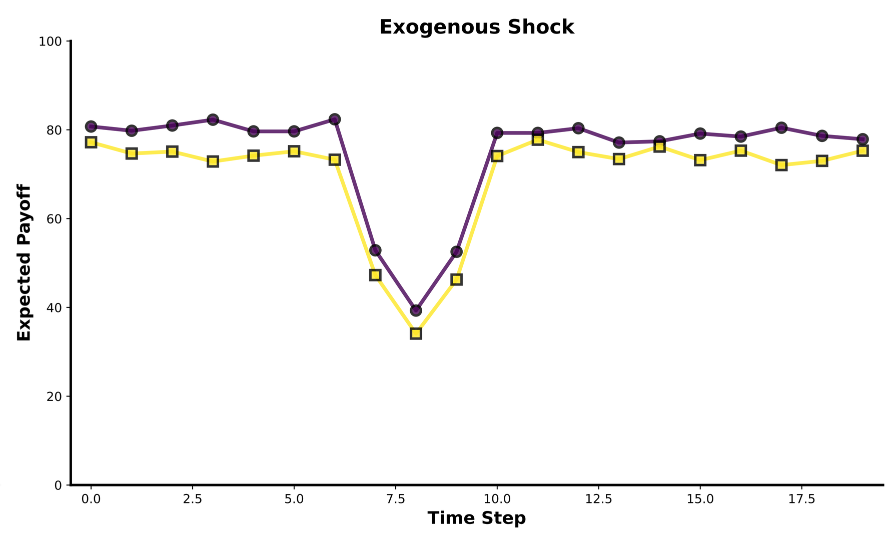
For t timesteps:
1. Pick action $A_t \in \mathcal{A}$,
2. Receive reward $X_t \sim \nu_{A_t}$ sampled i.i.d.
3. Continue to next timestep or **stop** and recommend action $\hat{a}$.

We write $\mu_a \coloneqq \mathbb{E}[\nu_a]$ as the **expected payoff** of action a.

- The best action a^* is such that $\mu_{a^*} > \mu_a$
- We focus on the **Fixed confidence** setting, as opposed to the fixed time
- An algorithm is **δ-correct** if it returns the best action with probability at least $1-\delta$
- The goal is to keep the **expected stopping time** τ as small as possible

This approach assumes i.i.d. data, **in the real world data is often non-stationary.**





Shifting Means

- We allow **the reward distributions $\nu_{t,a}$ to shift adversarially between timesteps.**
- But the **best action always stays the best action** $\mu_{t,a^*} > \mu_{t,a}$ for all $a \neq a^*$.

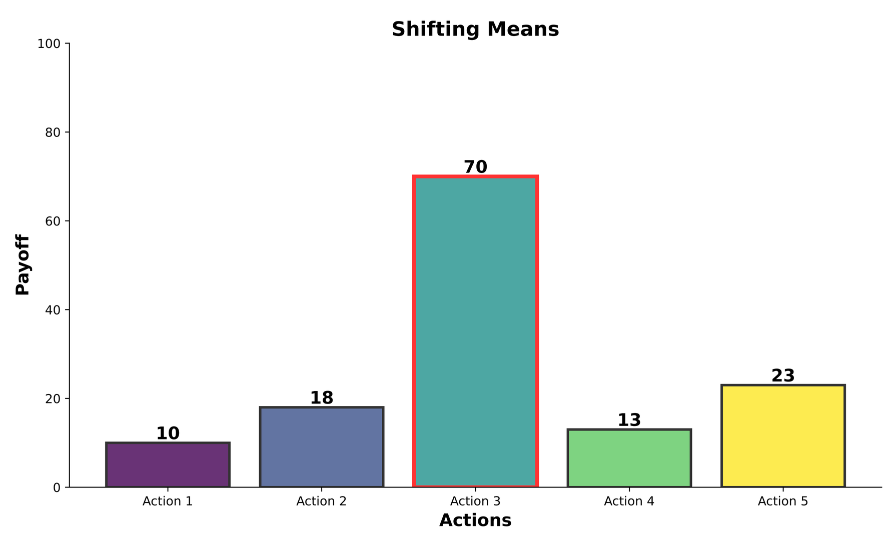
Best Arm Identification for Shifting Means

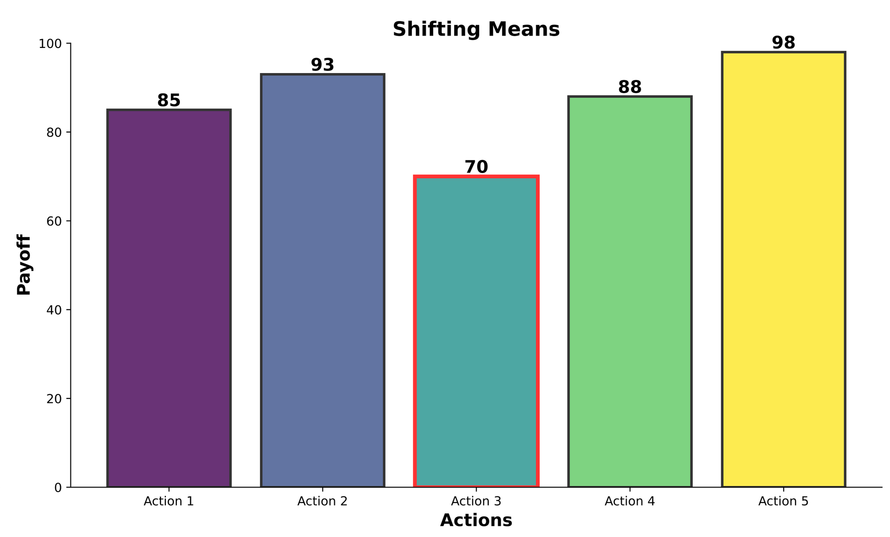
For t timesteps

1. Pick action $A_t \in \mathcal{A}$
2. Receive reward $X_t \sim \nu_{t,A_t}$
3. Continue to next timestep or **stop** and recommend action $\hat{a}$

- We define the expected payoff for each action $\mu_{t,a} \coloneqq \mathbb{E}[\nu_{t,a}]$.
- We assume that the samples X_t are still independent.

We only observe a single sample each timestep, how can we tell these scenarios apart?





Additional Assumptions & Notation:

- $\nu_{t,a}$ are sub-gaussian with variance proxy of at most σ^2
- Expected rewards are bounded $|\mu_{t,a}| \leq U$ (but X_t can be unbounded)
- If gaps are constant, we define Δ_{min} as the smallest gap, $\Delta_{\text{min}} \coloneqq \arg\min_{b \neq a} \mu_{t,a} - \mu_{t,b}$

The Generalized Likelihood Ration Test and How It Fails

The Generalized Likelihood Ration Test (**GLRT**) is a core part of the Track & Stop framework (Garivier and Kaufmann, 2016), which has lead to optimal rates in a wide variate of BAI settings.

- We interact with bandit μ, which has two actions a, b. There are two scenarios:
  - a is better than b, $\mu_{t,a} > \mu_{t,b}$ for all t **OR** a is not better than b, $\mu_{t,a} \leq \mu_{t,b}$ for all t

$$Z^{\text{GLRT}}(t) \coloneqq \log \frac{\max\limits_{\mu'_{t,a} > \mu'_{t,b}} \prod_{s=1}^{t} \mathbb{P}(X_s \mid \mu'_{s,A_s})}{\max\limits_{\lambda_{t,a} \leq \lambda_{t,b}} \prod_{s=1}^{t} \mathbb{P}(X_s \mid \lambda_{s,A_s})}$$

Pick $\mu'_{s,A_s} = X_s$ to maximize likelihood, and pick $\mu'_{s,\bar{A}_s}$ to fulfill the constraints; that is an optimal solution to either optimization problem.
$\Rightarrow Z^{\text{GLRT}}(t) = 0$ **for all t with probability** 1
We have t observations but $K \cdot t$ parameters and **the GLRT overfits.**

References

Garivier, A. and Kaufmann, E. (2016). Optimal best arm identification with fixed confidence. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 998–1027, Columbia University, New York, New York, USA. PMLR.

SMUS

Require: Action set $\mathcal{A}$, confidence δ
1: **for** $t = 1, \ldots$: **do**
2: Sample $A_t \sim \text{Unif}(\mathcal{A})$
3: Compute $Z_{a,b}(t)$ and c_t as in Equations 1 and 2 respectively
4: **if** there exists an $a \in \mathcal{A}$ such that $\min_{b \in \mathcal{A}, b \neq a} Z_{a,b}(t) \geq c_t$ **then**
5: **return** $\hat{a} = a$
6: **end if**
7: **end for**

Algorithm 1: SMUS (Shifting Means with Uniform Sampling)

With the current analysis, the best thing to do is uniform sampling.

Estimating Gaps

This is almost adversarial bandits $\Rightarrow$ We can estimate $\mu_{t,a}$ and $\mu_{t,b}$ with **importance weighting**.

- We define the gap $\Delta_{a,b}(t)$ between actions a, b at timestep t

$$\Delta_{a,b}(t) \coloneqq \mu_{t,a} - \mu_{t,b} \qquad \text{and it's estimator} \qquad \hat{\Delta}_{a,b}(t) \coloneqq \frac{\mathbb{I}[A_t = a]X_t}{w_t(A_t)} - \frac{\mathbb{I}[A_t = b]X_t}{w_t(A_t)} ,$$

where $w_t(A_t)$ is the probability that the algorithm picks action A_t at timestep t conditioned on the history.

- We define our test statistic for the evidence that action a is better than action b as

$$Z_{a,b}(t) \coloneqq \sum_{s=1}^{t} \frac{\mathbb{I}[A_s = a]X_s}{w_s(A_s)} - \frac{\mathbb{I}[A_s = b]X_s}{w_s(A_s)} \tag{1}$$

If $Z_{a,b}(t)$ grows large, then we know that a is better than b with growing at an expected rate of

$$\mathbb{E}\left[Z_{a,b}(t)\right] = \sum_{s=1}^{t} \Delta_{a,b}(s) .$$

- In general the gaps can vanish with t, for example $\Delta_{a,b}(t) = \frac{1}{t^2}$, then $\sum_{s=1}^{\infty} \Delta_{a,b}(s) = \frac{\pi^2}{6}$.
- We assume that $\sum_{s=1}^{t} \Delta_{a,b}(s) \geq \alpha t^{\beta}$.

Upper Bound - Theorem 4

Let $\nu_{t,a}$ be sub-gaussian with variance proxy of at most σ^2, $|\mu_{t,a}| \leq U$, and

$$c_t \coloneqq K\sqrt{2t\left(\sigma^2 + U^2\right)\log\left(K\delta^{-1}t^2\right)} , \tag{2}$$

*then **SMUS** is **δ-correct.***
*If $\beta = 1$, then **SMUS** has an expected stopping time of*

$$\mathbb{E}_{\mu}[\tau] \leq \frac{16K^2\left(\sigma^2 + U^2\right)}{\alpha^2} \log\left(\delta^{-1}\right) + \frac{32K^2\left(\sigma^2 + U^2\right)}{\alpha^2} \log\left(\frac{32K^2\left(\sigma^2 + U^2\right)}{\alpha^2}\right) ,$$

*and for $\beta \neq 1$ **check the paper**. Furthermore, if the gaps are constant, then $\alpha = \Delta_{\text{min}}$ and $\beta = 1$ and **SMUS** achieves*

SMUS on Constant Gaps

$$\mathbb{E}[\tau] \leq \frac{16K^2\left(\sigma^2 + U^2\right)}{\Delta_{\text{min}}^2} \log\left(\delta^{-1}\right) + \frac{32K^2\left(\sigma^2 + U^2\right)}{\Delta_{\text{min}}^2} \log\left(\frac{32K^2\left(\sigma^2 + U^2\right)}{\Delta_{\text{min}}^2}\right) + 4 .$$

General Proof Plan:
1. Characterize largest size $Z_{a,b}(t)$ can be when b is better than a,
   - $\Rightarrow$ Derive c_t and safety guarantee using **Ville's inequality**
2. Find point in time t_1 where **$Z_{a,b}(t)$ is larger than** c_t with high probability when a is better than b,
   - For that to happen at all, we require that $\beta > \frac{1}{2}$ as $c_t \in O(\sqrt{t \log t})$
   - Chernoff works here
3. Find $\mathbb{E}_{\mu}[\tau]$ by bounding everything before t_1 trivially and everything after t_1 carefully using our high probability bound.
   - The location of t_1 is almost all of the expected stopping time

Lower Bound

Let the gaps be constant (see the paper for varying gaps). If an algorithm is δ-correct on a group of bandits, then it is also δ-correct on any distribution over those bandits, allowing us to use **randomisation in our lower bound.**

Definition 5. *We say that random variable $B \in \mathbb{R}$ is a (C, R)-**good randomisation scheme** if*

- *$B \in [-R, R]$ with probability 1,*
- *For any $\Delta \in \mathbb{R}$ and zero-mean normal $\epsilon \sim \mathcal{N}(0, \sigma^2)$, we have*

$$\text{KL}(\epsilon + B \parallel \Delta + \epsilon + B) \leq \frac{\Delta^2}{2(\sigma^2 + C)} .$$

Theorem 7. *If there exists $(C, U/2)$-**good randomisation**, any δ-correct algorithm must suffer at least*

$$\mathbb{E}_{\mu}[\tau] \geq \frac{8(\sigma^2 + C)}{\Delta_{\min}^2} \text{kl}(\delta \parallel 1-\delta) .$$

With **uniform sampling** we can obtain a **$(\sigma U/2, U/2)$-good randomisation scheme.**

Our Lower Bound for Constant Gaps

$$\mathbb{E}_{\mu}[\tau] \in \Omega\left(\frac{\left(\sigma^2 + \sigma U\right)}{\Delta_{\min}^2} \log \delta^{-1}\right)$$

BAI (Garivier and Kaufmann, 2016)

$$\mathbb{E}_{\mu}[\tau] \in \Omega\left(\sum_{b \in \mathcal{A} \setminus \{a^*\}}^{K} \frac{\sigma^2}{\Delta_{a^*,b}^2} \log \delta^{-1}\right)$$

Gap of K^2U or KU^2 to the upper bound.

Lukas.Zierahn@gmail.com