Multi-Armed Bandits with Self-Information Rewards

Michal Yemini
michal.yemini@biu.ac.il

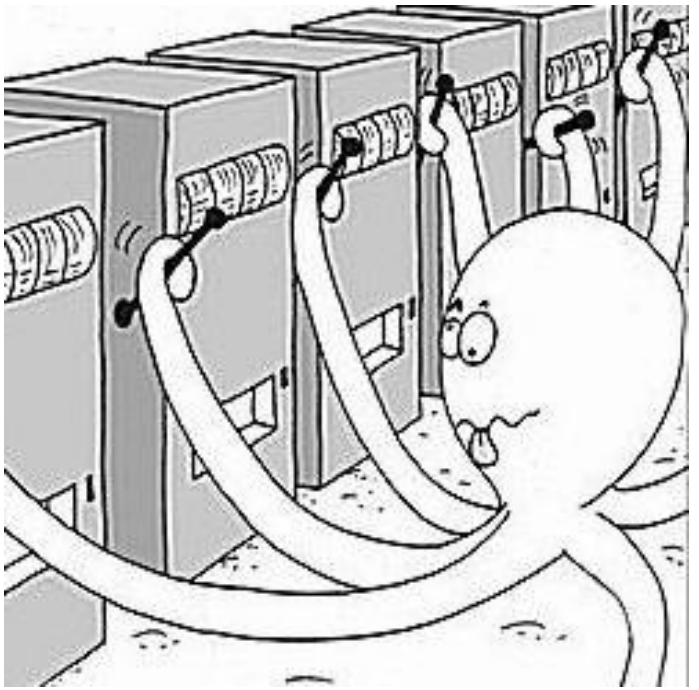
Joint work with Nir Weinberger

The Alexander Kofkin
Faculty of Engineering
Bar-Ilan University

August 28, 2023

Multi-Armed Bandits

- K arms with unknown rewards.
- At each time t a player chooses to play an arm $I(t)$ and receives a **reward** $X_{I(t),t}$.
- The **rewards** of each arm i are **i.i.d.** with mean μ_i.
- The player aims to maximize its total expected reward, or minimize its pseudo-regret



$$R(t) = t \max_{i \in 1,\dots,K} \mu_i - \sum_{i \in [K]} \mathbb{E}(N_i(t)) \cdot \mu_i,$$

where $N_i(t) = \sum_{\tau=1}^{t} \mathbb{1}[I_\tau = i]$.

A logarithmic regret is attainable with a UCB approach for the mean reward estimation.

Information as a Reward

- We consider a different reward structure, which is based on the **informativness** of the arm.
- A player has the goal of sampling from the most informative source.
- We focus on the natural choice of entropy with practical applications such as, coverage, ecosystem recovery and anomaly detection in mind.

Informational Multi-Armed Bandits

$\{X_i\}_{i=1}^K$ be memoryless sources, each defined on an alphabet $\mathcal{X}_i$.

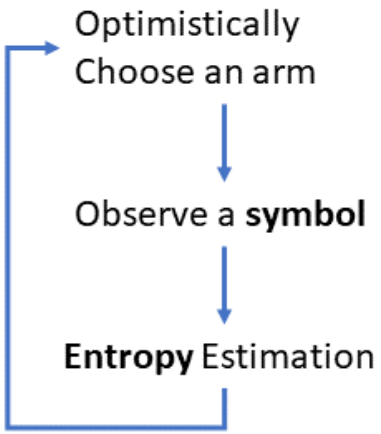
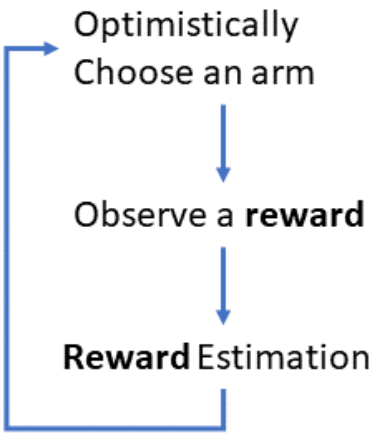
Denote $p_i(x) := \mathbb{P}[X_i = x]$ and let $p_i = \{p_i(x)\}_{x \in \mathcal{X}_i}$ the PMF of the ith source.

- At each round t, the player chooses one of the sources $i \in [K] := \{1, 2 \ldots, K\}$ and observes the **symbol** $X_i(t)$ from that source.
- The reward associated with this arm choice and this random observation is the **self-information:** $-\log p_i(X_i(t))$.
- The goal of the player is to choose the arm with the maximal expected reward, i.e., the maximal entropy, $i^* \in \operatorname*{argmax}_{i \in [K]} H_i$. This is equivalent to minimizing the **pseudo-regret:**

$$R(t) := t \cdot H_{i^*} - \sum_{i \in [K]} \mathbb{E}(N_i(t)) \cdot H_i. \tag{1}$$

The player does not know in advance the PMFs p_i, nor the entropy values H_i.

Informational MAB vs. Classical MAB

	Informational MAB	Classical MAB
Model flow		
Player's observation	Instantaneous **symbol**	Instantaneous **reward**
Reward estimation	**Biased** (Paninski 2003)	Unbiased

So far, UCB bounds have relied on the unbiasedness of the sampled mean of the rewards. **This does not hold in IMAB.**

A General UCB-Entropy Algorithm

Inputs: K, $\{\mathcal{X}_i\}_{i\in[K]}$, $\hat{H}(\cdot, n)$,$\mathrm{UCB}(\cdot,\cdot,n)$, α, $\delta_\alpha(t)$

set $\boldsymbol{X}_i(0) = \phi$ and $N_i(0) = 0$ for all $i \in [K]$

▷ Empty history sets at round $t = 0$

for $t = 1, 2, \ldots$ **do**

  play $I(t) \in \arg\max_{i\in[K]}\{\hat{H}(\boldsymbol{X}_i(t-1), N_i(t-1))) + \mathrm{UCB}(\boldsymbol{X}_i(t-1), \delta_\alpha(t), N_i(t-1))\}$

  set $\boldsymbol{X}_{I(t)}(t) = \boldsymbol{X}_{I(t)}(t-1) \cup X_{I(t)}(t)$ and $N_{I(t)}(t) = N_{I(t)}(t-1) + 1$

  ▷ Updating observation history of the current arm

  set $\boldsymbol{X}_i(t) = \boldsymbol{X}_i(t-1)$ and $N_i(t) = N_i(t-1)$ for all $i \in [K]\backslash I(t)$

  ▷ The observation set of other arms is unchanged

end for

return $\{N_i(t)\}_{i\in[K],\, t\in\mathbb{N}_+}$

▷ The number of times arm i was played by round t

How should we choose $\hat{H}(\boldsymbol{X}_i(t-1), N_i(t-1)))$ and $\mathrm{UCB}(\boldsymbol{X}_i(t-1), \delta_\alpha(t), N_i(t-1))$?

What is the resulting pseudo-regret?

UCB for Bias-Correction Entropy Estimation

Let p be a PMF defined on an alphabet $\mathcal{Y}$, and $\hat{p}(n)$ be the MLE p given n samples. Then

$$\hat{p}(y,n) := \frac{1}{n}\sum_{\ell=1}^{n} \mathbb{1}\{Y_\ell = y\}, \quad \text{for all } y \in \mathcal{Y}. \tag{2}$$

Upper bounding the MLE bias: (Paninski 2003)

$$H(p) - B(n) \leq \mathbb{E}[H(\hat{p}(n)) \leq H(p), \tag{3}$$

where $B(n) := \log\left(1 + \frac{|\mathcal{Y}|-1}{n}\right)$.

Thus, the bias-corrected term

$$\hat{H}(\hat{p}(n)) = H(\hat{p}(n)) + B(n) \tag{4}$$

has a **nonnegative bias**.

UCB for Bias-Correction Entropy Estimation (cont.)

UCB for bias correction estimator

Let $\boldsymbol{Y} = \{Y_\ell\}_{\ell \in [n]}$ be IID from a discrete distribution p over a *finite* alphabet $\mathcal{Y}$ such that $p(y) := \mathbb{P}[Y = y]$.
Then, assuming $n \geq 2$, it holds for any $\delta \in (0, 1)$ that

$$|H(\hat{p}(n)) + B(n) - H(p)| \leq \mathrm{UCB}_{\textit{bias}}(\delta, n), \tag{5}$$

with probability larger than $1 - \delta$, where

$$\mathrm{UCB}_{\mathsf{bias}}(\delta, n) := B(n) + \sqrt{\frac{2\log^2(n)}{n} \log\left(\frac{2}{\delta}\right)}. \tag{6}$$

This UCB leads to the following regret bound:

UCB for Bias-Correction Entropy Estimation (cont.)

Pseudo-regret for the bias-correction entropy estimation

Let $\Lambda_k(s) := s \cdot \log^k s$ and

$$\Gamma_{\text{bias}}(\alpha, \beta, \mathcal{Y}, \Delta, t) := \max\left\{ \frac{|\mathcal{Y}|-1}{e^{\frac{\beta\Delta}{2}} - 1}, 15\Lambda_2\left(\frac{8\log(2t^\alpha)}{(1-\beta)^2\Delta^2}\right)\right\}.$$

Assume that the general UCB-entropy algorithm is run with $\hat{H}(\boldsymbol{Y}, n) \equiv H(\hat{p}(n)) + B(n)$, and $\mathrm{UCB}(\boldsymbol{Y}, \delta, n) \equiv \mathrm{UCB}_{\textit{bias}}(\delta, n)$ with $\delta \equiv \delta_\alpha(t) = t^{-\alpha}$ and $\alpha > 2$. Let $\beta \in (0,1)$ be given. Then, the pseudo-regret is bounded as

$$R(t) \leq \sum_{i \in [K]:\Delta_i > 0} \left[\Gamma_{\textit{bias}}(\alpha, \beta, \mathcal{X}_i, \Delta_i, t) \cdot \Delta_i + \frac{2(\alpha-1)}{\alpha-2} \cdot \Delta_i \right].$$

Thus, the regret scales as $\tilde{O}(\frac{\log(t)}{\Delta_i})$, where the only difference from the standard UCB is the additional poly-logarithmic term.

UCB for Total-Variance Distance Based Entropy Estimation

- The bias term $B(n) := \log\left(1 + \frac{|\mathcal{Y}|-1}{n}\right)$ only depends on the alphabet size and is not sensitive to the source PMF.
- Can we improve the bounds whenever the entropy of sources is much less than the alphabet size?

Motivation:

Consider a Bernoulli arm for which $p_i(1) = \mathbb{P}[X_i = 1] \ll 1$.

The entropy of this arm is much smaller than the maximal possible value of $\log|\mathcal{X}_i| = \log(2)$.

A multiplicative Chernoff's inequality results in a confidence interval of $O(\sqrt{\frac{p_i(1)\log(1/\delta)}{n}})$ in the estimation of $p_i(1)$ using n samples from the source. Since $p_i(1) \ll 1$, this is much smaller than $O(\sqrt{\frac{\log(1/\delta)}{n}})$.

Nonetheless, $p_i(1)$ is unknown and thus must be estimated.
Hereafter, we focus on the binary IMAB.

UCB for Total-Variance Distance Based Entropy Estimation (cont.)

The proposed UCB algorithm and its regret analysis are based on the following inequality (that holds as long as $\mathrm{d_{TV}}(p,q) \leq \frac{1}{2}$):

$$|H(p) - H(q)| \leq \mathrm{d_{TV}}(p,q) \log\left(\frac{|\mathcal{Y}|}{\mathrm{d_{TV}}(p,q)}\right), \tag{7}$$

where $\mathrm{d_{TV}}(p,q) = \sum_{y\in\mathcal{Y}} |p(y) - q(y)|$ is the total variation distance between PMFs p and q defined on a common alphabet $\mathcal{Y}$.

UCB for Total-Variance Distance Based Entropy Estimation (cont.)

As a result we choose the following confidence interval function

$$\mathrm{UCB}_{\mathsf{ber}}(q,\delta,n) := \sqrt{\frac{12q\log\left(\frac{6}{\delta}\right)}{n}}\log\left(\frac{n}{q\log\left(\frac{6}{\delta}\right)}\right)+\frac{18\log\left(\frac{6}{\delta}\right)\log(n)}{n}, \tag{8}$$

and the corresponding confidence interval bound for the plug-in entropy estimator:

UCB for Binary Entropy Estimation

Let $\boldsymbol{Y}=\{Y_\ell\}_{\ell\in[n]}$ be IID from a Bernoulli with parameter $p=\mathbb{P}[Y_i=1]$, and let $\hat{p}(n)=\frac{1}{n}\sum_{\ell=1}^{n}\mathbb{1}\{Y_\ell=1\}$ be the empirical probability of $'1'$. Let $\delta\in[0,\frac{1}{2}]$ be given. If $n\geq 200\cdot\log\left(\frac{4}{\delta}\right)$ then

$$|h_b(\hat{p}(n))-h_b(p)|\leq \mathrm{UCB}_{\mathsf{ber}}(\hat{p}(n),\delta,n),$$

with probability larger than $1-\delta$.

UCB for Total-Variance Distance Based Entropy Estimation (cont.)

Pseudo-regret TV upper bound for binary sources

Assume that $\mathcal{X}_i = \{0,1\}$ for all $i \in [K]$ and that the general UCB-entropy algorithm is run with the plug-in entropy estimator $\hat{H}(\boldsymbol{Y},n) \equiv H(\hat{p}(\boldsymbol{Y},n))$ and upper confidence interval

$$\mathrm{UCB}(\boldsymbol{Y},\delta,n) \equiv \mathrm{UCB}_{ber}(\hat{p}(\boldsymbol{Y},n),\delta,n),$$

(as defined in (8)) with $\delta \equiv \delta_\alpha(t) = 6t^{-\alpha}$ with $\alpha > 2$. Then,

$$R(t) \leq \sum_{i\in[K]:\Delta_i>0} \inf_{\beta\in(0,1)} \Gamma_{ber}(\alpha,\beta,p_i(1),\Delta_i,t)\cdot\Delta_i + \frac{16(\alpha-1)}{\alpha-2}\cdot\Delta_i$$

where

$$\Gamma_{\mathrm{ber}}(\alpha,\beta,q,\Delta,t) := \max\left\{6\cdot\Lambda_1\left(\frac{36\alpha\log(t)}{(1-\beta)\Delta}\right),\right.$$
$$\left.\frac{5120q\alpha\log(t)}{\beta^2\Delta^2}\cdot\log^2\left(\frac{48}{\beta^2\Delta^2}\right), \frac{88\sqrt{\alpha\log(t)}}{\beta\Delta}\cdot\log\left(\frac{48}{\beta^2\Delta^2}\right)\right\},$$

Can we tighten this upper bound
on the pseudo-regret even further?

An asymptotic lower bound for binary IMABs

Let $\mathrm{D}_{\mathrm{KL}}(p||q) := p\log(p/q) + (1-p)\log((1-p)/(1-q))$ be the binary Kullback-Leibler divergence, where if $q \in \{0,1\}$ and $p \neq q$ then $\mathrm{D}_{\mathrm{KL}}(p||q) := \infty$.

LR lower bound

Consider the IMAB problem with K arms. A problem instance $\mathcal{I}$ is the collection $\{p_i\}_{i\in[K]}$ with $p_i \equiv p_i(1) \in [0, 1/2)$. Suppose that a IMAB algorithm is such that $R(t) = O(C_{\mathcal{I},a}t^a)$ for each problem instance I and $a > 0$. Then, for any instance $\mathcal{I}$,

$$\liminf_{t\to\infty} \frac{R(t)}{\log(t)} \geq \sum_{i\in[K]:\Delta_i>0} \frac{\Delta_i}{\mathrm{D}_{\mathrm{KL}}(p_i||p_{i^*})},$$

where $\Delta_i = \max_{j\in[K]} h_b(p_j) - h_b(p_i)$.

An asymptotic lower bound for binary IMABs (cont.)

Consider the case where $K = 2$ (two arms) and the two regimes:

- $p_1(1) = p$, $p_2(1) = p - \Lambda$ and $\Lambda \downarrow 0$. Then $\Delta = h_b(p_1) - h_b(p_2) = \Theta(\Lambda)$, $\mathrm{D}_{\mathrm{KL}}(p_2||p_1) = \Theta(\Lambda^2)$, and the ratio in the lower bound is $\Theta(\frac{\log t}{\Lambda})$. This is the upper bound we achieve using TV distance based entropy without the poly-logarithmic terms.
- $p_1(1), p_2(2) \approx \frac{1}{2}$, the binary entropy function "flattens", and is markedly different from the standard linear reward function. Assume that $p_1 = \frac{1}{2} - \Lambda$ and $p_2 = \frac{1}{2} - 2\Lambda$. Then, $\Delta = h_b(p_1) - h_b(p_2) = \Theta(\Lambda^2)$ and $\mathrm{D}_{\mathrm{KL}}(p_2||p_1) = \Theta(\Lambda^2)$, and the lower bound is asymptotically $\Theta(\log t)$ even if $\Lambda \downarrow 0$ and so also $\Delta \downarrow 0$. Thus, in this case we should explore ways to decrease the upper bound on the pseudo-regret.

Denote

$$\mathrm{UCB}_{\mathsf{ber}}^{(1/2)}(q,\delta,n) := 7\left|\frac{1}{2}-q\right|\cdot\sqrt{\frac{\log\left(\frac{4}{\delta}\right)}{n}}+\frac{9\log\left(\frac{4}{\delta}\right)}{n}. \tag{9}$$

Pseudo-regret TV upper bound for binary sources $p \approx 1/2$

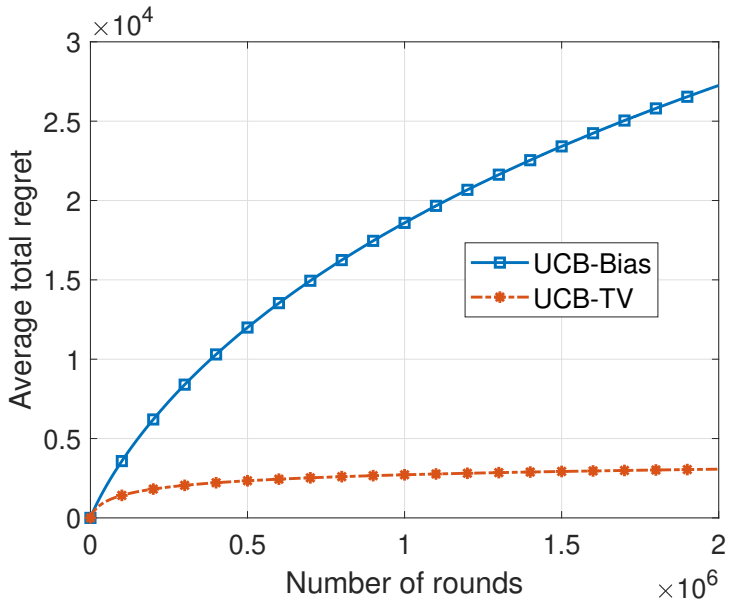
Let $\mathcal{X}_i = \{0,1\}$ and $\frac{2}{5} \leq p_2(1) < p_1(1) < \frac{1}{2}$ with $\Delta = h_b(p_1) - h_b(p_2)$. Further let the general UCB-entropy algorithm run with the plug-in entropy estimator $\hat{H}(\boldsymbol{Y},n) \equiv H(\hat{p}(\boldsymbol{Y},n))$ and $\mathrm{UCB}(\boldsymbol{Y},\delta,n) \equiv \mathrm{UCB}_{\mathit{ber}}^{(1/2)}(\hat{p}(\boldsymbol{Y},n),\delta,n)$ with $\delta \equiv \delta_\alpha(t) = 4t^{-\alpha}$ with $\alpha > 2$. Then,

$$R(t) \leq \frac{784\left(\frac{1}{2}-p_2(1)\right)^2\alpha\log(t)}{\Delta}+60\alpha\log(t)+\frac{8(\alpha-1)}{\alpha-2}\cdot\Delta.$$

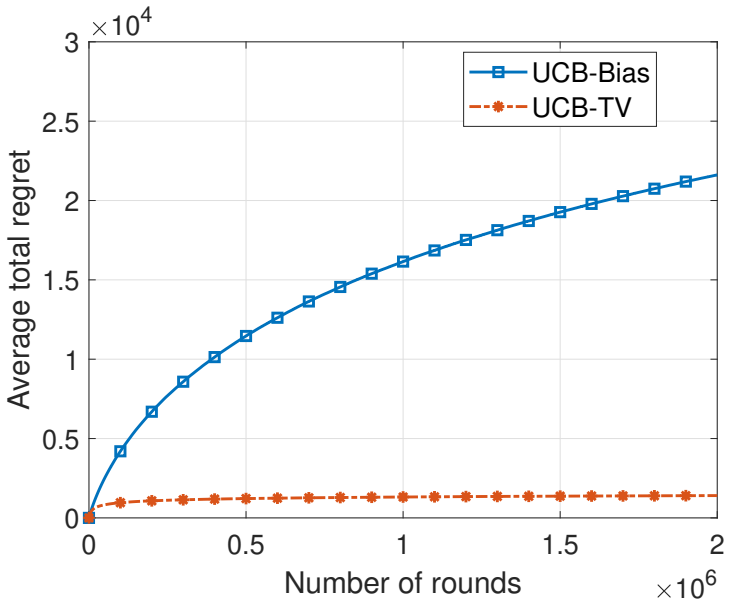
Sketch of the proof: approximate the binary entropy by its Taylor approximation. (In this case, $\left(\frac{1}{2}-p_2(1)\right)^2 = \Lambda^2 = \Theta(\Delta)$.)

Numerical Results - Binary Alphabet

Setup: two-armed IMAB with binary alphabets.



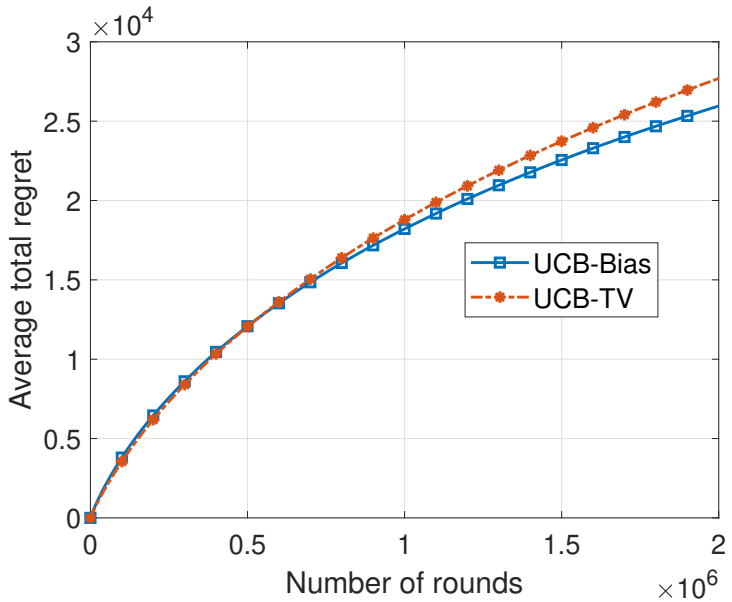
(a) $p_1(1) = 0.025$, and $p_2(1) = 10^{-4}$



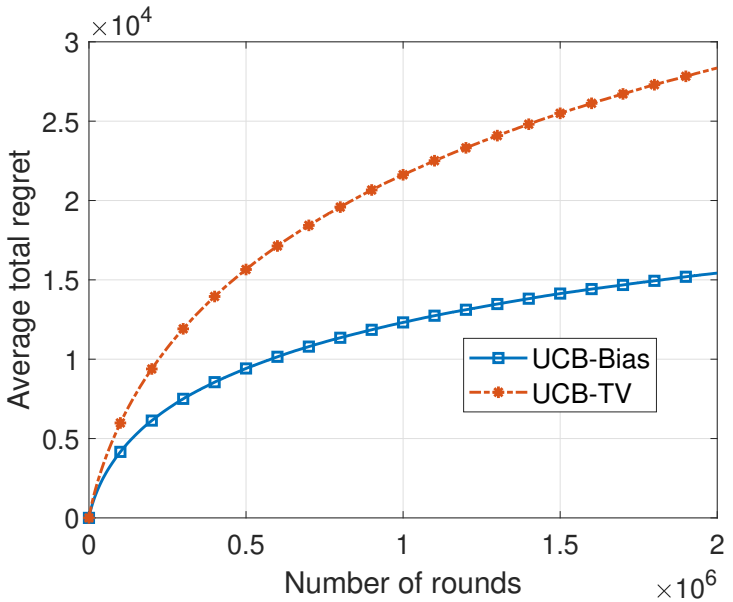
(b) $p_1(1) = 0.3$ and $p_2(1) = 0.15$

Numerical Results - Ternary Alphabet

Setup: two-armed IMAB with ternary alphabets.



(a) $p_1(0) = p_1(1) = 0.0125$, $p_2(0) = p_2(1) = 5 \times 10^{-5}$



(b) $p_1(0) = p_1(1) = 0.15$, $p_2(0) = p_2(1) = 0.075$

Open problems

Open problem 1: Conjecture

There exists an **entropy estimator** and a UCB for the IMAB that achieve the asymptotic lower bound for the binary case.
(Note that in this special case we can simply look for the arm with the smallest distance $|p_i(1) - 0.5|$, however, we are aiming to find schemes that can be extended (interpretable) to the general alphabet case.)

Open problem 2

We can extend the results of the binary alphabet to larger alphabets, however, the pseudo-regret bounds are not as tight. Is it possible to derive tighter upper bounds on the pseudo regret for a larger alphabet size?

Open problem 3

Explore the relation of the IMAB to heavy-tailed MAB.

Conclusions

- We introduced the informational multi-armed bandit problem with entropy rewards.
- The reward is not directly observed and its estimators suffer from bias.
- We presented a general UCB-entropy algorithm.
- We proposed two methods for estimating the entropy of the source and accompanied them with UCB.
- Numerical results.
- Open problems.