Stochastic & adversarial best-arm identification,

can we achieve the *best of both worlds?*

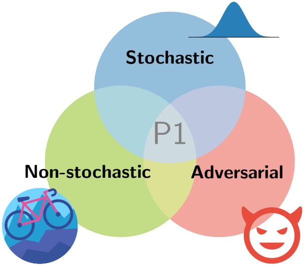
Victor Gabillon

joint work with Yasin Abbasi-Yadkori, Peter Bartlett,
Alan Malek & Michal Valko

Workshop on Quantifying Uncertainty: Stochastic, Adversarial, & Beyond,
Simons Institute, 13 September 2022

Summary

- **Setting:** A pure exploration bandit problem
- **Question:** Can one algorithm achieve BOB: Perform well under data-generating regimes either stochastic () or non-stochastic (or even against an adversary) ?
- **Contributions:**
  - a study of the problem against
  - an impossibility result on the BOB question
  - a simple algorithm P1 for BOB matching the lower bound.



After an **exploration phase** of T pulls, a Learner tries to **identify** the arm with highest cumulative reward out of K arms.

Bandit feedback: The learner only observes the reward/gain of the arm it chooses to explore.

For $t = 1, 2, \ldots, T$,

- simultaneously, Learner picks arm $I_t \in [K]$, **(K arms)**
- Adversary / environment picks gain $\boldsymbol{g}_t \in [0, g_{max}]^K$.
- Then, the Learner observes $\boldsymbol{g}_{I_t,t}$.

Finally, Learner **recommends** an arm denoted $\boxed{1_T}$, hoping $\boxed{1_T} = \boxed{1_T}$,

where $\boxed{1_t} \triangleq \arg\max_{k\in[K]} G_{k,t}$ & $G_{k,t} = \frac{1}{t}\sum_{t'=1}^{t} \boldsymbol{g}_{k,t'}$

Indices 1 2 3 … K

Indices 1 2 3 4

G

Indices 1 2 3 4

G	20\$	10\$	5\$	0\$
				
Indices	3	1	2	4

G	20$	10$	5$	0$
				
Indices	3	1	2	4
True ranks	1	2	3	4

G	20$	10$	5$	0$

Indices	3= $\boxed{1}$	1= $\boxed{2}$	2= $\boxed{3}$	4= $\boxed{4}$
True ranks	1	2	3	4

$\boxed{k}$ is the **index** of the arm ranked k-th according to G.
i.e. $G_{\boxed{1}} > G_{\boxed{2}} \geq G_{\boxed{3}} \geq \ldots \geq G_{\boxed{k}} \geq \ldots \geq G_{\boxed{K}}$

G	20$	10$	5$	0$

Indices	3= $\boxed{1}$	1= $\boxed{2}$	2= $\boxed{3}$	4= $\boxed{4}$
True ranks	1= 3	2= 1	3= 2	4= 4

$\boxed{k}$ is the **index** of the arm ranked k-th according to G.
i.e. $G_{\boxed{1}} > G_{\boxed{2}} \geq G_{\boxed{3}} \geq \ldots \geq G_{\boxed{k}} \geq \ldots \geq G_{\boxed{K}}$

k is the **rank** of the arm (of index) k according to G.

$\boxed{k}$ is the **index** of the arm ranked k-th w.r.t. to an estimate of $G_\cdot$: $\widehat{G}_\cdot, \widetilde{G}_\cdot$.

i.e. $G_{\boxed{1}} > G_{\boxed{2}} \geq G_{\boxed{3}} \geq \ldots \geq G_{\boxed{k}} \geq \ldots \geq G_{\boxed{K}}$

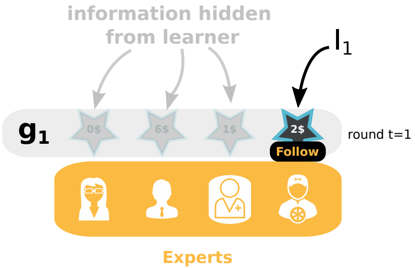
k is the **rank** of the arm (of index) k w.r.t. to an estimate of $G_\cdot$: $\widehat{G}_\cdot, \widetilde{G}_\cdot$.

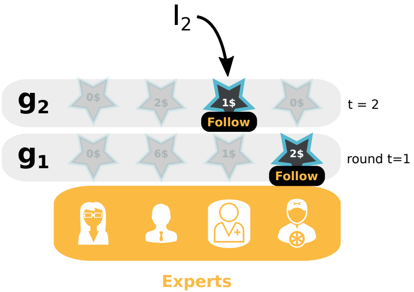

1
2
K=4

Products

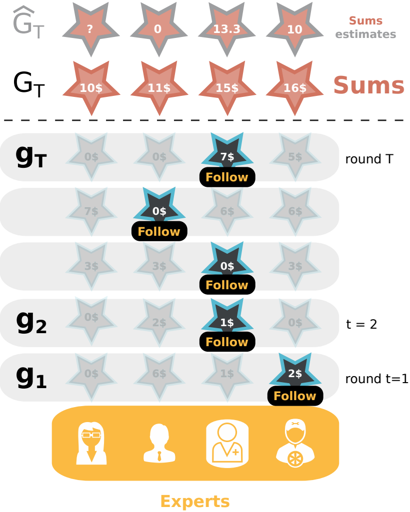
Drugs


Experts

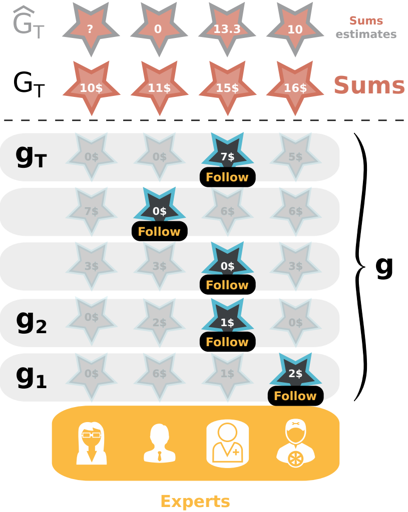

information hidden
from learner
I_1
g_1
0$
6$
1$
2$
Follow
round t=1
Experts

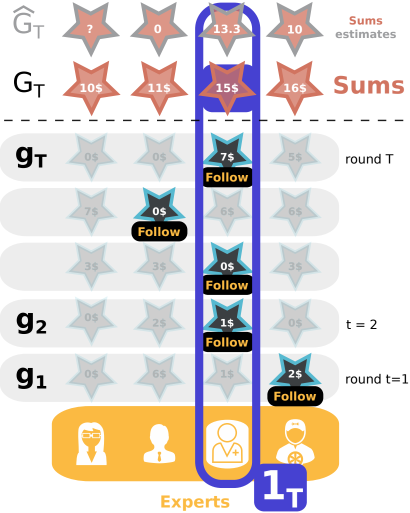


I_2
g_2
0$
2$
1$
Follow
0$
t = 2
g_1
0$
6$
1$
2$
Follow
round t=1
Experts

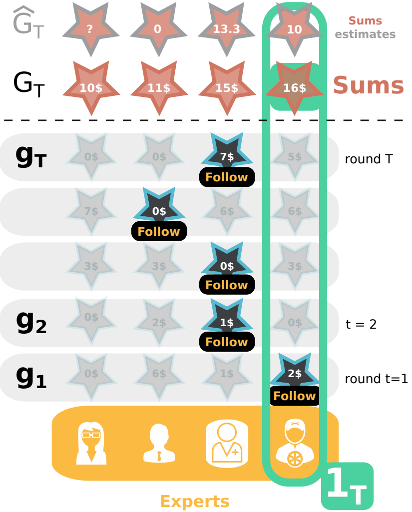


$\widehat{G}_T$
?
0
13.3
10
Sums estimates
G_T
10$
11$
15$
16$
Sums
g_T
0$
0$
7$
5$
Follow
round T
7$
0$
6$
6$
Follow
3$
3$
0$
3$
Follow
g_2
0$
2$
1$
0$
Follow
t = 2
g_1
0$
6$
1$
2$
Follow
round t=1
Experts


$\widehat{G}_T$
?
0
13.3
10
Sums estimates
G_T
10$
11$
15$
16$
Sums
g_T
0$
0$
7$
Follow
5$
7$
0$
Follow
6$
6$
3$
3$
0$
Follow
3$
g_2
0$
2$
1$
Follow
0$
g_1
0$
6$
1$
2$
Follow
g
Experts



$\widehat{G}_T$
?
0
13.3
10
Sums estimates
G_T
10$
11$
15$
16$
Sums
g_T
0$
0$
7$
Follow
5$
round T
7$
0$
Follow
6$
6$
3$
3$
0$
Follow
3$
g_2
0$
2$
1$
Follow
0$
t = 2
g_1
0$
6$
1$
2$
Follow
round t=1
1_T
Experts


$\widehat{G}_T$
?
0
13.3
10
Sums
estimates
G_T
10$
11$
15$
16$
Sums
g_T
0$
0$
7$
Follow
5$
round T
7$
0$
Follow
6$
6$
3$
3$
0$
Follow
3$
g_2
0$
2$
1$
Follow
0$
t = 2
g_1
0$
6$
1$
2$
Follow
round t=1
1_T
Experts

Cumulative regret

- $\mathrm{R}(T) = max_k(G_k) - \sum_{t=1}^{T} g_{I_t,t} = \text{16\$} - \left(\text{2\$} + \text{1\$} + \cdots + \text{7\$}\right)$
- Minimize the cumulative regret $\Leftrightarrow$ Play 1_T as often as possible
- **Exploration vs Exploitation**
- **Classic algorithms:** Thompson Sampling, UCB

Probability of misidentification — simple regret

- $e(T) = \mathbb{P}\left(1_T \neq 1_T\right)$ or $r(T) = G_{1_T} - G_{1_T} = \text{16\$} - \text{15\$}$
- Minimize the simple regret $\Leftrightarrow$ Identify 1_T
- **Pure Exploration**
- **Classic algorithms:** Hoeffding Race, Successive Rejects

Reward generation

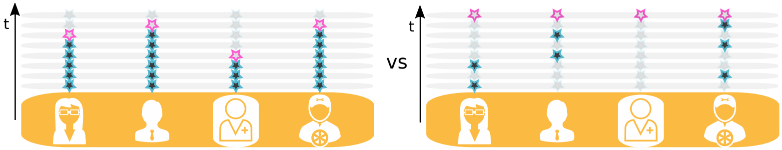
How are the rewards, $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_T$, generated?

Stochastic	Non stochastic	Adversarial
$\boldsymbol{g}_{k,t} \overset{\text{iid}}{\sim} \nu_k$, mean μ_k	Drop iid, non-stationary	Arbitrary $\boldsymbol{g} = \{\boldsymbol{g}_{k,t}\}_{k\in[K],t\in[T]}$
$1_T = \arg\max_{k\in[K]} \mu_k$		$1_T = \arg\max_{k\in[K]} G_k$
		$G_k = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{g}_{k,t}$
indifferent to $e(T)$		picks $\boldsymbol{g}$ maximizing $e(T)$
Related works:		
Hoefdding race [1]	[4][5][6]	New in Best arm identification
Successive Rejects (SR) [2]	(more on the next slide!)	Similar to adversarial bandit [3]

[1]: Maron & Moore, 1993, [2]: Audibert, Bubeck & Munos, 2010, [3]: Auer, Cesa-Bianchi, Freund & Schapire, 2002, [4]: Jamieson & Talwalkar, 2016, [5]: Allesiardo, Féraud & Maillard, 2017 , [6]: Altschuler, Brunel & Malek, 2019

Related works in non-stochastic best arm identification

- **Jamieson & Talwalkar, 2016** for hyperparameter optimization:
  - $\boldsymbol{g}_{k,t}$ are fixed by an adversary with the condition that $\boldsymbol{g}_{k,t}$ converge as $h_k = \lim_{t \to +\infty} \boldsymbol{g}_{k,t}$ exists.
  - At round t for its m-th pull of arm k, their learner observe $\boldsymbol{g}_{k,m}$, whereas our learner observes $\boldsymbol{g}_{k,t}$ (less hidden information).



- **Allesiardo, Féraud & Maillard, 2017**:
  $\boldsymbol{g}_{k,t}$ are sampled from a non-stationary process with the condition that the identity of the best arm so far does not change with time: $\mathbf{1}_t = \mathbf{1}_{t'}$, $\forall (t, t') \in [T]^2$.

- Corruption/contamination, **Altschuler, Brunel & Malek, 2019**:
  $\boldsymbol{g}_{k,t}$ are sampled i.i.d. but the learner observes $\boldsymbol{g}_{k,t} + c_{k,t}$ where $c_{k,t}$ can be an arbitrary corruption.

- **Deterministic Uniform exploration** (DETER-UNIFORM).
  Pull every arm deterministically T/K times.

- **Successive Rejects** (SR) **(Audibert, Bubeck & Munos, 2010)**
  Pull more the arms with highest estimated average reward.

- **Deterministic Uniform exploration** (DETER-UNIFORM).
  Pull every arm deterministically T/K times.

- **Successive Rejects** (SR) **(Audibert, Bubeck & Munos, 2010)**
  Pull more the arms with highest estimated average reward.

The estimated mean of arm k at time t is simply the standard average:

$$\boxed{\widehat{\mu}_{k,t}} \triangleq \frac{\sum_{t'=1}^{t} \mathbf{1}\{I_{t'} = k\} g_{k,t'}}{\sum_{t'=1}^{T} \mathbf{1}\{I_{t'} = k\}}$$

- **Deterministic Uniform exploration** (DETER-UNIFORM).
  Pull every arm deterministically T/K times.

- **Successive Rejects** (SR) **(Audibert, Bubeck & Munos, 2010)**
  Pull more the arms with highest estimated average reward.

The estimated mean of arm k at time t is simply the standard average:

$$\widehat{\mu}_{k,t} \triangleq \frac{\sum_{t'=1}^{t} \mathbf{1}\{I_{t'} = k\} g_{k,t'}}{\sum_{t'=1}^{T} \mathbf{1}\{I_{t'} = k\}}$$

- **UCB-Exploration** **(Audibert, Bubeck & Munos, 2010)**
  Pull $\arg\max_{k\in[K]} \widehat{\mu}_{k,t} + g_{max}\sqrt{\frac{a}{T_k}}$, $a \in \mathbb{R}$, T_k: # of pulls of k.

$g_{max}\sqrt{\frac{a}{T_k}}$ is the *uncertainty* on $\widehat{\mu}_{k,t}$ but requires knowledge of g_{max}.

- **Deterministic Uniform exploration** (DETER-UNIFORM).
  Pull every arm deterministically T/K times.

- **Successive Rejects** (SR) **(Audibert, Bubeck & Munos, 2010)**
  Pull more the arms with highest estimated average reward.

The estimated mean of arm k at time t is simply the standard average:

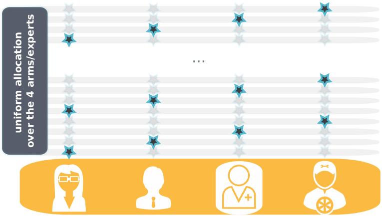
$$\boxed{\widehat{\mu}_{k,t}} \triangleq \frac{\sum_{t'=1}^{t} \mathbf{1}\{I_{t'} = k\} g_{k,t'}}{\sum_{t'=1}^{T} \mathbf{1}\{I_{t'} = k\}}$$

	$e_{\blacktriangle}(T)$	$e_{\text{😈}}(T)$
DETER-UNIFORM	✗	?
SR	✓	?

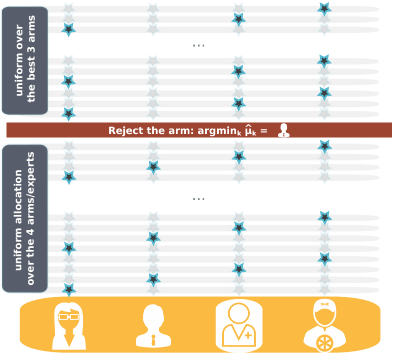
- SR is an elimination algorithm pulling uniformly over a set of remaining candidate arms.
- The arm k, ranked -th by SR, is allocated T/ pulls

- SR is an elimination algorithm pulling uniformly over a set of remaining candidate arms.

- The arm k, ranked -th by SR, is allocated T/ pulls



- SR is an elimination algorithm pulling uniformly over a set of remaining candidate arms.
- The arm k, ranked -th by SR, is allocated T/ pulls

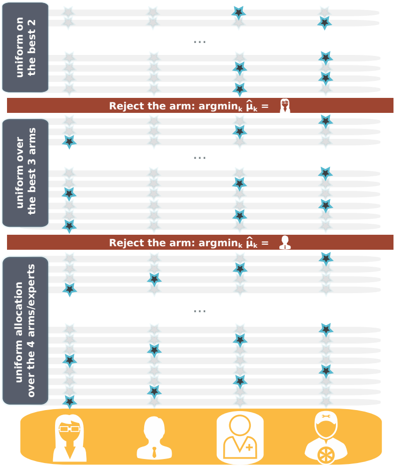


The estimated mean of arm k at time t is simply the standard average:

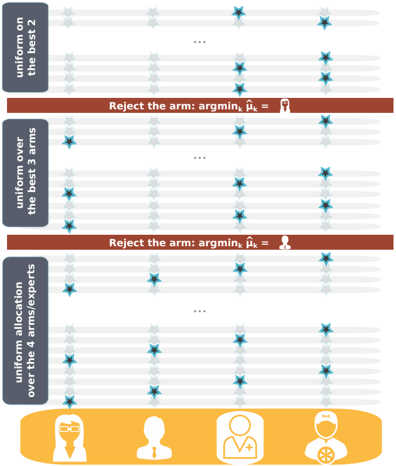
$$\boxed{\widehat{\mu}_{k,t}} \triangleq \frac{\sum_{t'=1}^{t} \mathbf{1}\{I_{t'}=k\} \boldsymbol{g}_{k,t'}}{\sum_{t'=1}^{T} \mathbf{1}\{I_{t'}=k\}}$$

- SR is an elimination algorithm pulling uniformly over a set of remaining candidate arms.

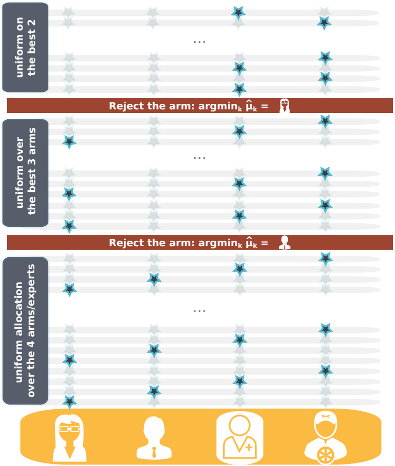
- The arm k, ranked -th by SR, is allocated T/ pulls

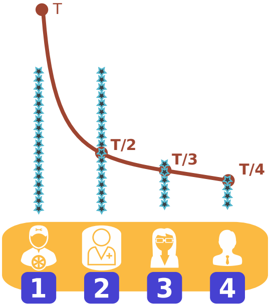


- SR is an elimination algorithm pulling uniformly over a set of remaining candidate arms.
- The arm k, ranked k-th by SR, is allocated T/k pulls



- SR is an elimination algorithm pulling uniformly over a set of remaining candidate arms.
- The arm k, ranked k-th by SR, is allocated T/k pulls

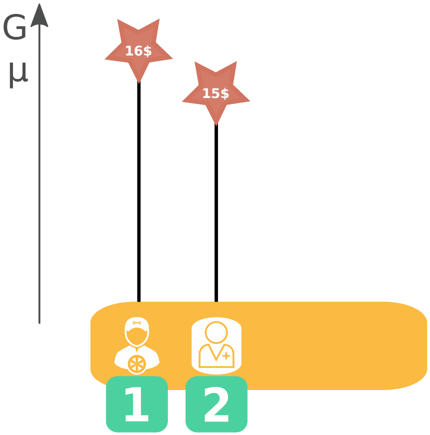


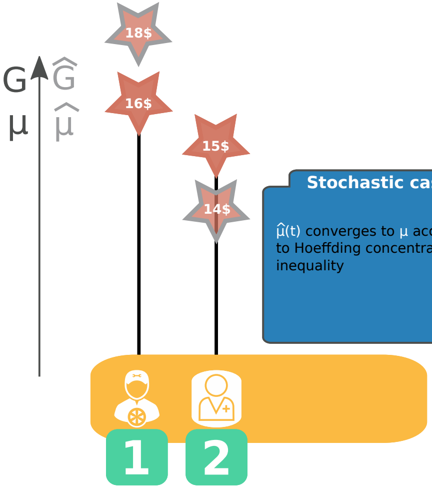


	$e_{\blacktriangle}(T)$	$e_{😈}(T)$
DETER-UNIFORM	✗ ? ?	?
SR	✓ ? ?	?

And now...

Let us precise the $e_{\blacktriangle}(T)$ for the uniform and SR algorithms.

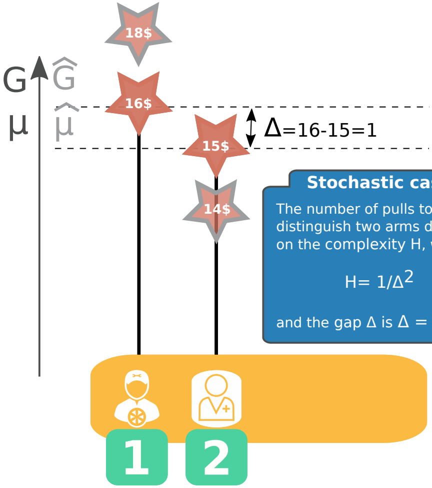

G
μ
16$
15$
1
2



Stochastic case

$\widehat{\mu}(t)$ converges to μ according to Hoeffding concentration inequality

Gaps and complexities in hindsight

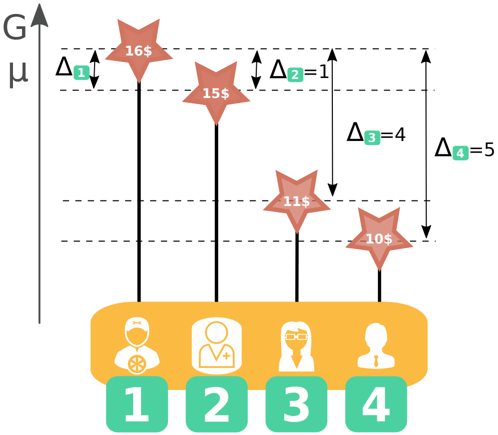


Stochastic case

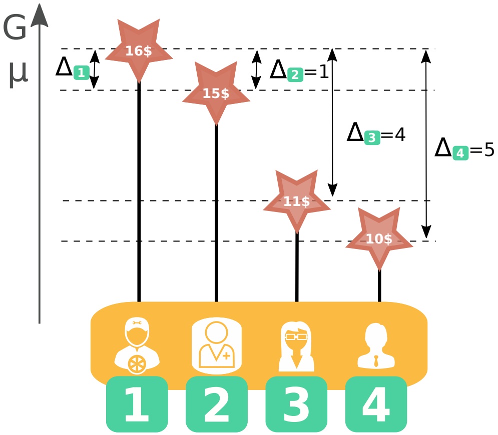
The number of pulls to distinguish two arms depends on the complexity H, where:

$$H = 1/\Delta^2$$

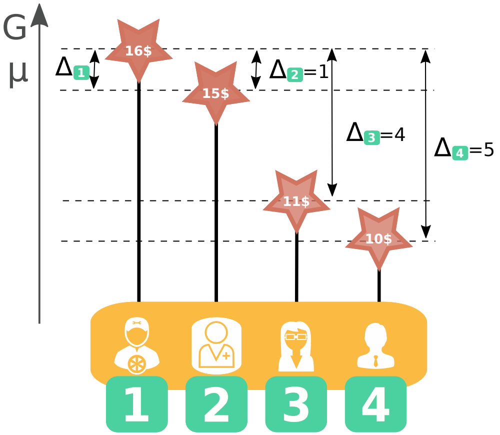
and the gap Δ is $\Delta = \mu_1 - \mu_2$



To distinguish arm k from arm 1, the learner must have its *uncertainty* on μ_k (or G_k) smaller than Δ_k, i.e. $|\widehat{\mu}_k - \mu_k| \leq \Delta_k/2$.

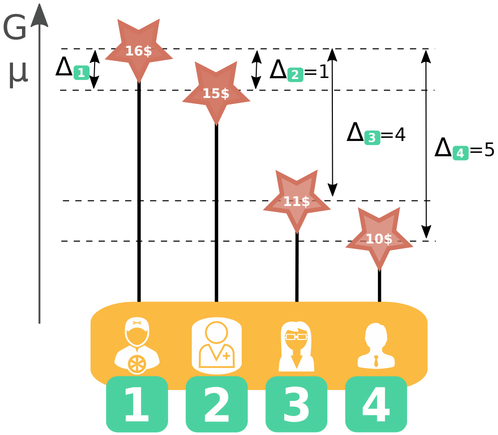


$$H_{\mathrm{UNIF}} \triangleq \frac{K}{\Delta_{1}^{2}}$$



$$H_{\text{UNIF}} \triangleq \frac{K}{\Delta_{1}^{2}} \qquad \& \qquad H_{\text{SR}} \triangleq \max_{k\in[K]} \frac{k}{\Delta_{k}^{2}}$$

Stochastic case



$$H_{\text{UNIF}} \triangleq \frac{K}{\Delta_{1}^{2}} \qquad \geq \qquad H_{\text{SR}} \triangleq \max_{k\in[K]} \frac{k}{\Delta_{k}^{2}} = \widetilde{\mathcal{O}}\left(\sum_{k\in[K]} \frac{1}{\Delta_{k}^{2}}\right)$$

Stochastic case

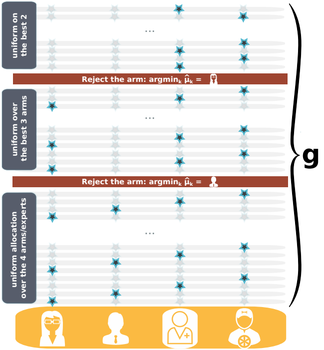
		$e_{\blacktriangle}(T)$	$e_{😈}(T)$
DETER-UNIFORM	✗	$e^{\frac{-T}{H_{\textbf{UNIF}}}}$	?
SR	✓	$e^{\frac{-T}{H_{\textbf{SR}} \log K}}$	?

And now...

Let us discuss SR against an adversary.

SR can be tricked by an adversary arranging $\boldsymbol{g}$

- SR pulls the arm deterministically (will hide rewards easily)
- SR stops pulling arms (reject) during the game (hides rewards)
- SR uses the standard estimation of the average $\widehat{\mu}_{k,t}$ (**biased** against)



- The learner needs to use internal randomization
- The learner should be careful about rejecting arm: no rejection!
- Be careful of the bias of the reward estimates.

		$e(T)$		$e(T)$
DETER-UNIFORM	✗	$e^{\frac{-T}{H_{\text{UNIF}}}}$	✗	1
SR	✓	$e^{\frac{-T}{H_{\text{SR}} \log K}}$	✗	1
? ? ? ?			✓	

The $(e_{\blacktriangle}, e_{😈})$ table so far

		$e_{\blacktriangle}(T)$		$e_{😈}(T)$
DETER-UNIFORM	✘	$e^{\frac{-T}{H_{\mathsf{UNIF}}}}$	✘	1
SR	✔	$e^{\frac{-T}{H_{\mathsf{SR}} \log K}}$	✘	1
? ? ? ?			✔	

And now...

Let us discuss the adversarial 😈 setting

•DETER-UNIFORM•:

- Pull every arm deterministically T/K times.
- Recommend the arm with highest $\widehat{\mu}_{k,t}$

Robutifying

- **Internal randomization**: pull arm k at time t with proba $p_{k,t} = \mathbb{P}(I_t = k)$
- Replace $\widehat{\mu}_{k,t}$ by $\widetilde{G}_{k,t}$ as $\mathrm{E}[\widetilde{G}_{k,t}] = G_{k,t}$ **(unbiased)**

$$\widehat{\mu}_{k,t} \triangleq \frac{\sum_{t'=1}^{t} \mathbf{1}\{I_{t'} = k\} g_{k,t'}}{\sum_{t'=1}^{T} \mathbf{1}\{I_{t'} = k\}} \qquad \widetilde{G}_{k,t} = \frac{1}{t} \sum_{t'=1}^{t} \frac{g_{k,t'}}{p_{k,t'}} \mathbf{1}\{I_{t'} = k\}$$

(importance weights)

•RULE•:

- At time t, pull arm k with probability $p_{k,t} = 1/K$.
- Recommend the arm with highest $\widetilde{G}_{k,t}$

Theorem (RULE vs. 😈)

For any T and adversarial $\boldsymbol{g}$, RULE *satisfies*

$$e_{😈}(T) = \mathcal{O}\left(\exp\left(-\frac{T}{H_{\mathrm{UNIF}(g)}}\right)\right).$$

The proof uses a Bernstein bound.

Theorem (Rule vs. 😈)

For any T and adversarial $\boldsymbol{g}$, Rule *satisfies*

$$e_{\text{😈}}(T) = \mathcal{O}\left(\exp\left(-\frac{T}{H_{\mathsf{UNIF}(g)}}\right)\right).$$

The proof uses a Bernstein bound.

Theorem (😈 Lower bound)

For any learner, a $\boldsymbol{g}$ of complexity H_{UNIF},

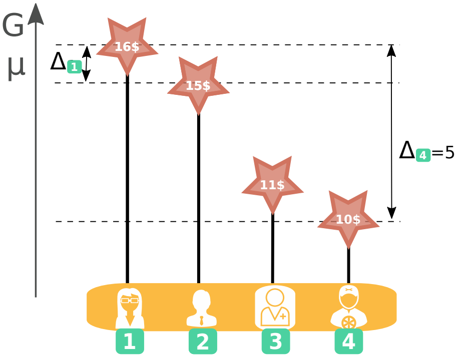
$$e_{\text{😈}}(T) = \Omega\left(\exp\left(-\frac{T}{H_{\mathsf{UNIF(g)}}}\right)\right).$$

Rule: optimal gap-dependent rates against 😈.

Idea: The can force the learner to have, at $t = T/2$, an uncertainty on $\widetilde{G}_{k,T/2}$ of order Δ_{1}, $\forall k \in [K]$ (instead of the usual Δ_{k} in).

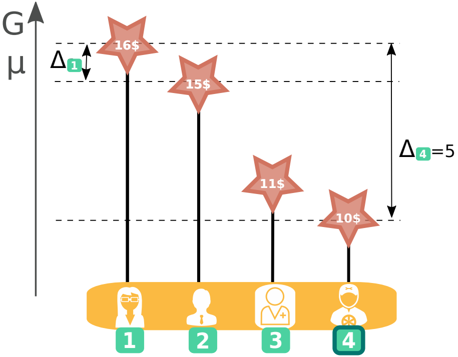
Our proof of the lower bound uses some arguments of **Audibert & Bubeck (2010)**, **Carpentier and Locatelli (2016)** and **Auer and Chiang (2016)**

Idea: The can force the learner to have, at $t = T/2$, an uncertainty on $\widetilde{G}_{k,T/2}$ of order Δ_{1}, $\forall k \in [K]$ (instead of the usual Δ_{k} in).



Given a Learner and a bandit PROBLEM I defined for the first half of the game (until $t = T/2$)

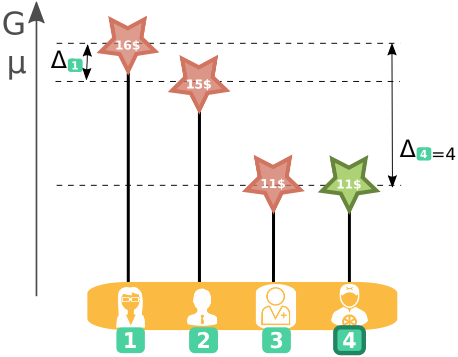
Idea: The can force the learner to have, at $t = T/2$, an uncertainty on $\widetilde{G}_{k,T/2}$ of order Δ_{1}, $\forall k \in [K]$ (instead of the usual Δ_{k} in).



At least one arm is pulled less than $T/(2K)$ by the Learner (not pulled enough to detect small variations of size Δ_{1}, of its mean $\leftarrow$ prone to error). Here its arm 

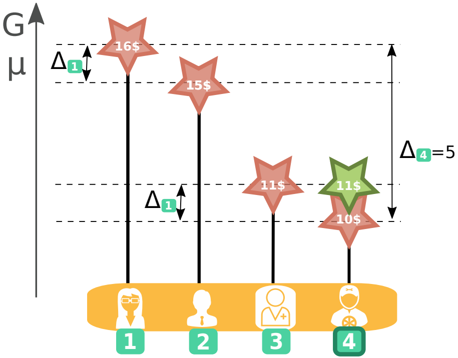
Proof sketch of the lower bound against

Idea: The can force the learner to have, at $t = T/2$, an uncertainty on $\widetilde{G}_{k,T/2}$ of order Δ_{1}, $\forall k \in [K]$ (instead of the usual Δ_{k} in).



Then, an alternative/similar PROBLEM II is created, by modifying by Δ_{1}.
PROBLEM II is defined for $t = 1$ to $t = T/2$.

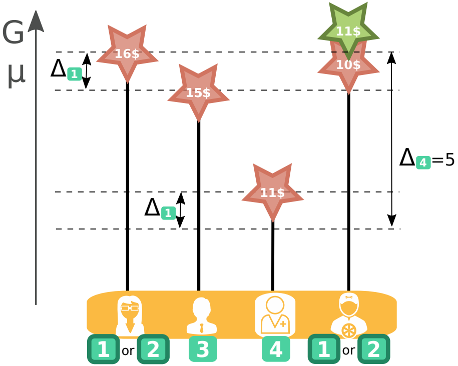
Idea: The can force the learner to have, at $t = T/2$, an uncertainty on $\widetilde{G}_{k,T/2}$ of order Δ_{1}, $\forall k \in [K]$ (instead of the usual Δ_{k} in).



This is the superposition of PROBLEM I & II.

PROBLEM I & II are indistinguishable with proba $e^{-\frac{T\Delta_{1}^{2}}{K}}$

Idea: The can force the learner to have, at $t = T/2$, an uncertainty on $\widetilde{G}_{k,T/2}$ of order Δ_{1}, $\forall k \in [K]$ (instead of the usual Δ_{k} in).

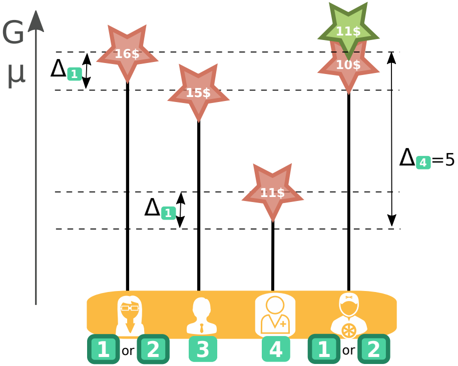


Between $t = T/2$ and $t = T$, the arm get enough reward so that:

$=$ 1 in PROBLEM II while $=$ 1 in PROBLEM I

Proof sketch of the lower bound against

Idea: The can force the learner to have, at $t = T/2$, an uncertainty on $\boxed{\widetilde{G}_{k,T/2}}$ of order Δ_{1}, $\forall k \in [K]$ (instead of the usual Δ_{k} in).



The lower bound comes from the fact that PROBLEM I & II

- have different best arms
- are indistinguishable w.p. $e^{\frac{-T\Delta_{1}^2}{K}}$, i.e. $\underbrace{P_{II}\left(1 = \right)}_{\text{error in II}} \geq \underbrace{P_{I}\left(1 = \right)}_{\text{success in I}} e^{\frac{-T\Delta_{1}^2}{K}}$

- Best arm identification against is too hard: uniform exploration (RULE) is optimal.
- However RULE is suboptimal in .
- SR, optimal in fails against

	$e(T)$		$e(T)$	
SR	✔	$e^{\frac{-T}{H_{\text{SR}} \log K}}$	✘	1
RULE	✘	$e^{\frac{-T}{H_{\text{UNIF}}}}$	✔	$e^{\frac{-T}{H_{\text{UNIF}}}}$

The question:

Is there a learner performing optimally in **both** *the stochastic and adversarial cases while not being aware of the nature of the rewards*

	$e(T)$	$e(T)$
	$e^{\frac{-T}{H_{\text{SR}} \log K}}$	$e^{\frac{-T}{H_{\text{UNIF}}}}$

- Best arm identification against 😈 is too hard: uniform exploration (Rule) is optimal.
- However Rule is suboptimal in .
- SR, optimal in fails against 😈

	$e(T)$		$e_{😈}(T)$	
SR	✔	$e^{\frac{-T}{H_{\mathrm{SR}} \log K}}$	✘	1
Rule	✘	$e^{\frac{-T}{H_{\mathrm{UNIF}}}}$	✔	$e^{\frac{-T}{H_{\mathrm{UNIF}}}}$

The BOB question:

¿ *Is there a learner performing optimally in* **both** *the stochastic and adversarial cases while not being aware of the nature of the rewards* ?

	$e(T)$		$e_{😈}(T)$	
? ? ? ?	✔	$e^{\frac{-T}{H_{\mathrm{SR}} \log K}}$	✔	$e^{\frac{-T}{H_{\mathrm{UNIF}}}}$

- Best arm identification against is too hard: uniform exploration (RULE) is optimal.
- However RULE is suboptimal in .
- SR, optimal in fails against

	$e(T)$		$e(T)$	
SR	✓	$e^{\frac{-T}{H_{\mathsf{SR}} \log K}}$	✗	1
RULE	✗	$e^{\frac{-T}{H_{\mathsf{UNIF}}}}$	✓	$e^{\frac{-T}{H_{\mathsf{UNIF}}}}$

The BOB question:

¿ *Is there a learner performing optimally in* **both** *the stochastic and adversarial cases while not being aware of the nature of the rewards* ?

The BOB question was studied in the cumulative regret setting in **Bubeck & Slivkins, 2012, Seldin & Slivkins, 2014, Auer & Chiang, 2016, Zimmert & Seldin, 2018...**

New notion of complexity

$$H_{\mathrm{BOB}} \triangleq \frac{1}{\Delta_{1}} \max_{k \in [K]} \frac{k}{\Delta_{k}}.$$

Theorem (Lower bound for the BOB challenge)

For any learner,

if *for all adversarial problem* $\boldsymbol{g}$,

$$e(T) \leq \frac{1}{16},$$

then *there exists a stochastic problem with complexity* H_{BOB} *such that*

$$e(T) \geq \frac{1}{64} \exp\left(-\frac{2048\,T}{H_{\mathrm{BOB}}}\right) \overset{\text{sometimes}}{=} \frac{1}{64} \exp\left(-\frac{2048\,T}{H_{\mathrm{SR}}\sqrt{K}}\right).$$

New notion of complexity

$$H_{\mathrm{BOB}} \triangleq \frac{1}{\Delta_{1}} \max_{k\in[K]} \frac{k}{\Delta_{k}}.$$

Theorem (Lower bound for the BOB challenge)

For any learner,

if *for all adversarial problem* $\boldsymbol{g}$,

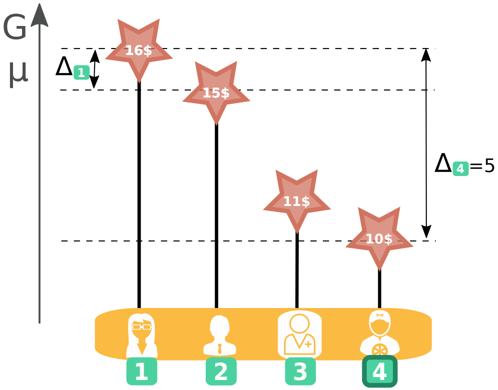
$$e(T) \leq \frac{1}{16},$$

then *there exists a stochastic problem with complexity* H_{BOB} *such that*

$$e(T) \geq \frac{1}{64}\exp\left(-\frac{2048\,T}{H_{\mathrm{BOB}}}\right) \overset{\text{sometimes}}{=} \frac{1}{64}\exp\left(-\frac{2048\,T}{H_{\mathrm{SR}}\sqrt{K}}\right).$$

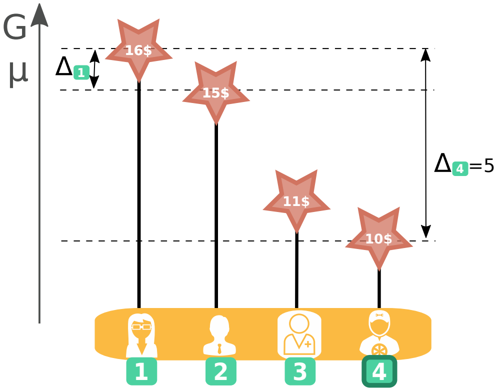
Idea: The 😈 can force the learner to have, $\forall k \in [K]$, at $t = T\frac{\Delta_{1}}{\Delta_{k}}$(instead of $t = T$ in ▲), an uncertainty on $\widetilde{G}_{k,t}$ of order Δ_{k}.

Idea: The 😈 can force the learner to have, $\forall k \in [K]$, at $t = T\frac{\Delta_1}{\Delta_k}$ (instead of $t = T$ in), an uncertainty on $\widetilde{G}_{k,t}$ of order Δ_k.



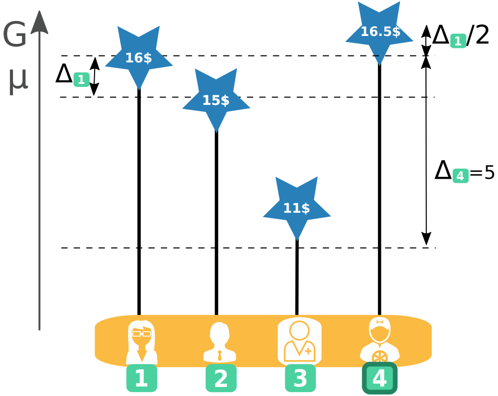
Given a Learner, a rank k, and a BASE PROBLEM defined until $t = T\frac{\Delta_1}{\Delta_k}$

Idea: The can force the learner to have, $\forall k \in [K]$, at $t = T\frac{\Delta_1}{\Delta_k}$ (instead of $t = T$ in), an uncertainty on $\widetilde{G}_{k,t}$ of order Δ_k.



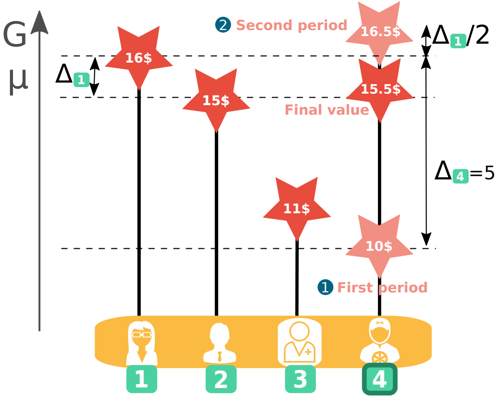
At least one arm, k', with $\Delta_{k'} \leq \Delta_k$, is pulled less than $\frac{T}{k}\frac{\Delta_1}{\Delta_k}$ (not pulled enough). Let us illustrate $k = 4$ and assume for simplicity $k' = 4 =$

Idea: The can force the learner to have, $\forall k \in [K]$, at $t = T\frac{\Delta_1}{\Delta_k}$ (instead of $t = T$ in), an uncertainty on $\widetilde{G}_{k,t}$ of order Δ_k.



A similar PROBLEM **STO** is created, , by modifying by $\Delta_4 + \Delta_1/2$.
PROBLEM **STO** is defined for $t = 1$ to $t = T$.

Idea: The 😈 can force the learner to have, $\forall k \in [K]$, at $t = T^{\frac{\Delta_1}{\Delta_k}}$ (instead of $t = T$ in ▲), an uncertainty on $\widetilde{G}_{k,t}$ of order Δ_k.



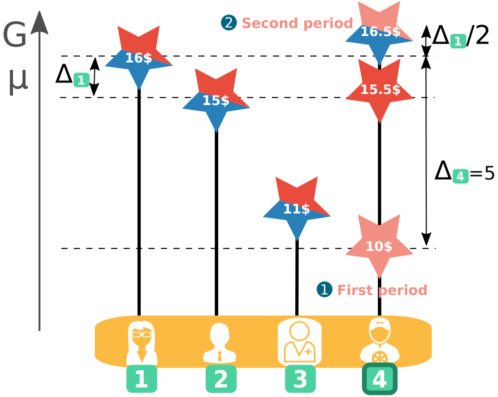
PROBLEM **ADV** 😈: ❶ follows BASE from $t = 1$ to $t = T^{\frac{\Delta_1}{\Delta_k}}$,

❷ follows **STO** afterwards .

Modifying of Δ_k during $T^{\frac{\Delta_1}{\Delta_k}}$ changes the means of Δ_1 w.r.t **STO**.

Proof sketch of the BOB lower bound

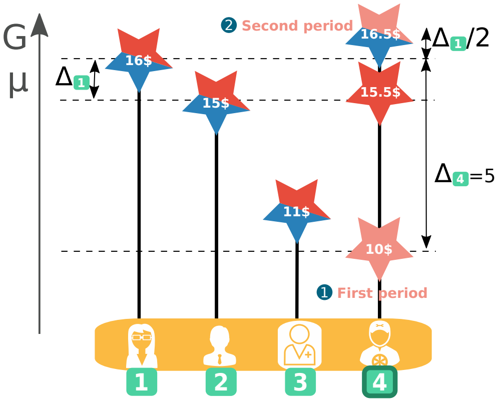
Idea: The can force the learner to have, $\forall k \in [K]$, at $t = T\frac{\Delta_1}{\Delta_k}$ (instead of $t = T$ in), an uncertainty on $\widetilde{G}_{k,t}$ of order Δ_k.



Superposition of **ADV** & **STO**: = 1 in **STO** and = 1 in **ADV**,

ADV & **STO** are indistinguishable w. p. $e^{-\frac{\left(T\frac{\Delta_1}{\Delta_k}\right)\Delta_k^2}{k}} = e^{-\frac{T\Delta_1\Delta_k}{k}}$

Idea: The can force the learner to have, $\forall k \in [K]$, at $t = T\frac{\Delta_1}{\Delta_k}$ (instead of $t = T$ in), an uncertainty on $\widetilde{G}_{k,t}$ of order Δ_k.



The lower bound comes from, a **max** over $k \in [K]$, and the fact that **ADV** & **STO**

- have different best arms,
- are indistinguishable w.p. $e^{-\frac{T\Delta_1\Delta_k}{k}}$: $\underbrace{P_{\text{STO}}\left(1 = \square\right)}_{\text{error in II}} \geq \underbrace{P_{\text{ADV}}\left(1 = \square\right)}_{\text{success in I}} e^{-\frac{T\Delta_1\Delta_k}{k}}$

New notion of complexity

$$H_{\mathrm{BOB}} \triangleq \frac{1}{\Delta_{1}} \max_{k\in[K]} \frac{k}{\Delta_{k}} .$$

Theorem (Lower bound for the BOB challenge)

For any learner,

if *for all adversarial problems* $\boldsymbol{g}$,

$$e_{😈}(T) \leq \frac{1}{16},$$

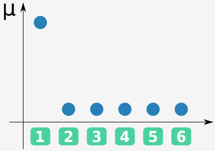
then *there exists a stochastic problem with complexity* H_{BOB} *such that*

$$e_{\blacktriangle}(T) \geq \frac{1}{64} \exp\left(-\frac{2048\,T}{H_{\mathrm{BOB}}}\right) \overset{\text{sometimes}}{=} \frac{1}{64} \exp\left(-\frac{2048\,T}{H_{\mathrm{SR}}\sqrt{K}}\right).$$

$$H_{\mathrm{SR}} \le H_{\mathrm{BOB}} \le H_{\mathrm{UNIF}}.$$

$$\underbrace{\max_{k\in[K]} \frac{k}{\Delta^2_{\boxed{k}}}}_{H_{\mathrm{SR}}} \le \underbrace{\frac{1}{\Delta_{\boxed{1}}} \max_{k\in[K]} \frac{k}{\Delta_{\boxed{k}}}}_{H_{\mathrm{BOB}}} \le \underbrace{\frac{1}{\Delta^2_{\boxed{1}}}}_{H_{\mathrm{UNIF}}}.$$

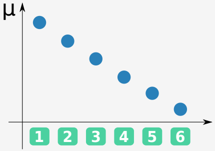
▶ Flat regime



$$H_{\mathrm{SR}} = H_{\mathrm{BOB}} = H_{\mathrm{UNIF}}$$

BOB is achieved by Rule.

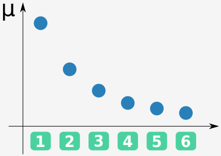
▶ Linear regime



$$H_{\mathrm{SR}} = H_{\mathrm{BOB}} = \frac{H_{\mathrm{UNIF}}}{K}$$

BOB can be achieved but not by RULE.
Need a new learner!

▶ Square-root regime



$$H_{\mathrm{SR}} = \frac{H_{\mathrm{BOB}}}{\sqrt{2K}} = \frac{H_{\mathrm{UNIF}}}{K}$$

No learner can do BOB!

There is still an open question!

The new BOB question:

¿ *Can an algorithm achieve the following* ?

	$e(T)$	$e(T)$
? ? ? ?	✓ $e^{\frac{-T}{H_{\text{BOB}} \log K}}$	✓ $e^{\frac{-T}{H_{\text{UNIF}}}}$

Why is the BOB question challenging?

- **Bias** of estimator $\boxed{\widehat{\mu}_{k,t}} = \frac{\sum_{t'=1}^{t} \mathbf{1}\{I_{t'}=k\} g_{k,t'}}{\sum_{t'=1}^{T} \mathbf{1}\{I_{t'}=k\}}$ (simple average)
- **Variance** of $\boxed{\widetilde{G}_{k,t}} = \sum_{t'=1}^{t} \frac{g_{k,t'}}{p_{k,t'}} \mathbf{1}\{I_{t'} = k\}$ (importance weights)

We use :

- Pulling uniformly for too long with $p_{k,t} = \frac{1}{K}$ leads to a large **variance**, up to being of order K, in $\widetilde{G}_{k,t}$.
- Objective: reduce the variance (**uncertainty**) of the estimators of the best arms ≈ find the best arm

The new BOB question:

Can an algorithm achieve the following?

	$e_{\blacktriangle}(T)$	$e_{😈}(T)$
? ? ? ?	✔ $e^{\frac{-T}{H_{\mathrm{BOB}} \log K}}$	✔ $e^{\frac{-T}{H_{\mathrm{UNIF}}}}$

Why is the BOB question challenging?

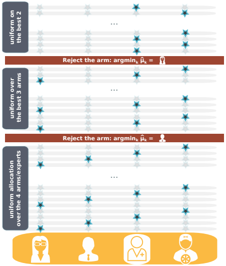
- Bias of estimator $\widehat{\mu}_{k,t} = \frac{\sum_{t'=1}^{t} \mathbf{1}\{I_{t'}=k\} g_{k,t'}}{\sum_{t'=1}^{T} \mathbf{1}\{I_{t'}=k\}}$ (simple average)
- Variance of $\widetilde{G}_{k,t} = \sum_{t'=1}^{t} \frac{g_{k,t'}}{p_{k,t'}} \mathbf{1}\{I_{t'} = k\}$ (importance weights)

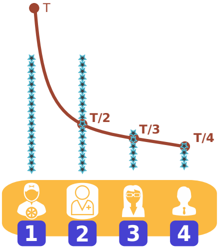
We use $\widetilde{G}_{k,t}$:

- Pulling uniformly for too long with $p_{k,t} = \frac{1}{K}$ leads to a large **variance**, up to being of order K, in $\widetilde{G}_{k,t}$.
- Objective: reduce the variance (**uncertainty**) of the estimators of the best arms $\approx$ find the best arm

Idea: Robustify the SR algorithm.

- We use $\widetilde{G}_{k,t}$
- Cannot pull uniformly, as in SR, for almost half of the game.
- Need to pull the estimated best arms earlier.
- Need to remove the rejections
- Reuse the proportions of SR (arm k, ranked k-th by SR, is allocated T/k pulls)





At any time t, $\textsc{P1}$ pulls

- arm **1** with 'probability' $\mathbf{1}$
- arm with 'probability' $\frac{1}{2}$
- arm with 'probability' $\frac{1}{3}$
- and so on...
- arm with 'probability' $\frac{1}{k}$
- and with 'probability' $\frac{1}{K}$
- (and normalize)

$\overline{\log}K = \sum_{k=1}^{K}(1/k)$, with $|\overline{\log}K - \log K| \leq 1$

$\textsc{P1}$ early bets are almost **costless**! (and necessary):

- The estimated best arms are prioritized since the first pull to reduce variance.
- Up to a $\log K$ factor, all arms are pulled uniformly.
- $\textsc{P1}$ implicitly control the uncertainty of the estimates.

At any time t, $\mathrm{P1}$ pulls

- arm **1** with 'probability' $\mathbf{1}$
- arm **2** with 'probability' $\mathbf{\frac{1}{2}}$
- arm with 'probability' $\frac{1}{3}$
- and so on...
- arm with 'probability' $\frac{1}{k}$
- and with 'probability' $\frac{1}{K}$
- (and normalize)

$\overline{\log}K = \sum_{k=1}^{K}(1/k)$, with $|\overline{\log}K - \log K| \leq 1$

$\mathrm{P1}$ early bets are almost **costless**! (and necessary):

- The estimated best arms are prioritized since the first pull to reduce variance.
- Up to a $\log K$ factor, all arms are pulled uniformly.
- $\mathrm{P1}$ implicitly control the uncertainty of the estimates.

At any time t, $\textsc{P1}$ pulls

- arm **1** with 'probability' $\mathbf{1}$
- arm **2** with 'probability' $\mathbf{\frac{1}{2}}$
- arm **3** with 'probability' $\mathbf{\frac{1}{3}}$
- and so on...
- arm with 'probability' $\frac{1}{k}$
- and with 'probability' $\frac{1}{K}$
- (and normalize)

$\overline{\log}K = \sum_{k=1}^{K}(1/k)$, with $|\overline{\log}K - \log K| \leq 1$

$\textsc{P1}$ early bets are almost **costless**! (and necessary):

- The estimated best arms are prioritized since the first pull to reduce variance.
- Up to a $\log K$ factor, all arms are pulled uniformly.
- $\textsc{P1}$ implicitly control the uncertainty of the estimates.

At any time t, P1 pulls

- arm **1** with 'probability' $\mathbf{1}$
- arm **2** with 'probability' $\frac{1}{2}$
- arm **3** with 'probability' $\frac{1}{3}$
- and so on...
- arm **k** with 'probability' $\frac{1}{k}$
- and **K** with 'probability' $\frac{1}{K}$
- (and normalize)

$\overline{\log}K = \sum_{k=1}^{K}(1/k)$, with $|\overline{\log}K - \log K| \leq 1$

P1 early bets are almost **costless**! (and necessary):

- The estimated best arms are prioritized since the first pull to reduce variance.
- Up to a $\log K$ factor, all arms are pulled uniformly.
- P1 implicitly control the uncertainty of the estimates.

At any time t, $\textsc{p}1$ pulls

- arm **1** with 'probability' $\mathbf{1}$
- arm **2** with 'probability' $\frac{1}{2}$
- arm **3** with 'probability' $\frac{1}{3}$
- and so on...
- arm **k** with 'probability' $\frac{1}{k}$
- and **K** with 'probability' $\frac{1}{K}$
- (and normalize)

$\overline{\log}K = \sum_{k=1}^{K}(1/k)$, with $|\overline{\log}K - \log K| \leq 1$

$\textsc{p}1$ early bets are almost **costless**! (and necessary):

- The estimated best arms are prioritized since the first pull to reduce variance.
- Up to a $\log K$ factor, all arms are pulled uniformly.
- $\textsc{p}1$ implicitly control the uncertainty of the estimates.

At any time t, $\mathrm{P1}$ pulls

- arm **1** with probability $1/\overline{\log}K$
- arm **2** with probability $\frac{1}{2\ \overline{\log}K}$
- arm **3** with probability $\frac{1}{3\ \overline{\log}K}$
- and so on...
- arm **k** with probability $\frac{1}{k\ \overline{\log}K}$
- and **K** with probability $\frac{1}{K\ \overline{\log}K}$
- (and normalize)

$\overline{\log}K = \sum_{k=1}^{K}(1/k)$, with $|\overline{\log}K - \log K| \leq 1$

$\mathrm{P1}$ early bets are almost **costless**! (and necessary):

- The estimated best arms are prioritized since the first pull to reduce variance.
- Up to a $\log K$ factor, all arms are pulled uniformly.
- $\mathrm{P1}$ implicitly control the uncertainty of the estimates.

At any time t, $\mathrm{P1}$ pulls

- arm $\boxed{1}$ with probability $1/\overline{\log}K$
- arm $\boxed{2}$ with probability $\frac{1}{2\ \overline{\log}K}$
- arm $\boxed{3}$ with probability $\frac{1}{3\ \overline{\log}K}$
- and so on...
- arm $\boxed{k}$ with probability $\frac{1}{k\ \overline{\log}K}$
- and $\boxed{K}$ with probability $\frac{1}{K\ \overline{\log}K}$
- (and normalize)

$\overline{\log}K = \sum_{k=1}^{K}(1/k)$, with $|\overline{\log}K - \log K| \leq 1$

$\mathrm{P1}$ early bets are almost **costless**! (and necessary):

- The estimated best arms are prioritized since the first pull to reduce variance.
- Up to a $\log K$ factor, all arms are pulled uniformly.
- $\mathrm{P1}$ implicitly control the uncertainty of the estimates.

For $t = 1, 2, \ldots$

- Rank the arms according to $\widetilde{G}_{k,t}$: Rank arm k as k_t.
- Select arm $I_t \in [K]$ with

$$\boldsymbol{p}_{k,t} \triangleq \mathbb{P}\left(I_t = k\right) \triangleq \frac{1}{k_t \, \overline{\log} K} \qquad \text{for all } k \in [K].$$

Recommend, at any round t, $1_t \triangleq \arg\max_{k \in [K]} \widetilde{G}_{k,t}$.

The algorithm is anytime and parameter-free.

Pseudo-code of P1

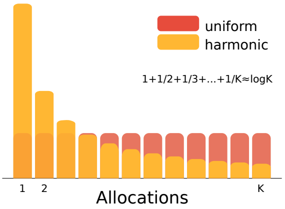
For $t = 1, 2, \ldots$

- Rank the arms according to $\widetilde{G}_{k,t}$: Rank arm k as k_t.
- Select arm $I_t \in [K]$ with

$$p_{k,t} \triangleq \mathbb{P}(I_t = k) \triangleq \frac{1}{k_t \, \overline{\log} K} \qquad \text{for all } k \in [K].$$

Recommend, at any round t, $1_t \triangleq \arg\max_{k \in [K]} \widetilde{G}_{k,t}$.

The algorithm is anytime and parameter-free.



Pseudo-code of P1

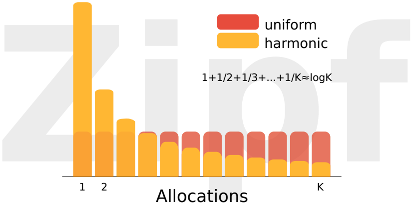
For $t = 1, 2, \dots$

- Rank the arms according to $\widetilde{G}_{k,t}$: Rank arm k as k_t.
- Select arm $I_t \in [K]$ with

$$\boldsymbol{p}_{k,t} \triangleq \mathbb{P}\left(I_t = k\right) \triangleq \frac{1}{k_t \, \overline{\log} K} \qquad \text{for all } k \in [K].$$

Recommend, at any round t, $1_t \triangleq \arg\max_{k \in [K]} \widetilde{G}_{k,t}$.

The algorithm is anytime and parameter-free.



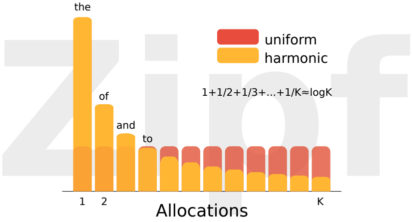
For $t = 1, 2, \ldots$

- Rank the arms according to $\widetilde{G}_{k,t}$: Rank arm k as k_t.
- Select arm $I_t \in [K]$ with

$$p_{k,t} \triangleq \mathbb{P}(I_t = k) \triangleq \frac{1}{k_t \, \overline{\log} K} \qquad \text{for all } k \in [K].$$

Recommend, at any round t, $1_t \triangleq \arg\max_{k \in [K]} \widetilde{G}_{k,t}$.

The algorithm is anytime and parameter-free.



Theorem (Upper bounds for P1)

For any problems:

- $$e_{\blacktriangle}(T) = \mathcal{O}\left(\exp\left(-\frac{T}{H_{BOB}\ \log^2(K)}\right)\right)$$
- $$e_{\text{😈}}(T) = \mathcal{O}\left(\exp\left(-\frac{T}{H_{\mathrm{UNIF}(g)}\ \overline{\log}(K)}\right)\right)$$

		$e_{\blacktriangle}(T)$		$e_{\text{😈}}(T)$
SR	✔	$e^{\frac{-T}{H_{\mathbf{SR}} \log K}}$	✘	1
RULE	✘	$e^{\frac{-T}{H_{\mathbf{UNIF}}}}$	✔	$e^{\frac{-T}{H_{\mathbf{UNIF}}}}$
P1	✔	$e^{\frac{-T}{H_{\mathbf{BOB}} \log K}}$	✔	$e^{\frac{-T}{H_{\mathbf{UNIF}}}}$

Early bets are costless / Early bets are necessary

- $\boldsymbol{p}_{k,t} \geq 1/(K\overline{\log}K)$ is enough to obtain the same complexity H_{UNIF} as RULE, up to a factor $\log K$.

- , $K-1$ arbitrary 'virtual' phases that each ends at round $T_i = Ta_i$. Chosen in hindsight to minimize the upper bound (P1 is oblivious to a_i).

Intuitively, after T_i the event ξ_i happens with high probability:

$$\xi_i \triangleq \left\{ \forall t > T_i, \forall k \in [K] : \mu_{1} - \mu_k < \frac{\Delta_i}{2} \implies k_t < i \right\}.$$

$\Rightarrow$ for any such arm k, for $t > T_i$, $\boldsymbol{p}_{k,t} \geq 1/(i-1)$.
$\Rightarrow$ smaller variance (of order $i-1$) in their estimates $\widetilde{\boldsymbol{g}}_{k,t}$
$\Rightarrow$ better estimates in the rest of the game.

The proof works iteratively over the phases.

Error $= \xi_i$ does not hold.

Trade off in setting the length of the phases with a_i:
Trade off between event ξ_i happening fast and ξ_i happening with high probability
Short phases $=$ not enough samples to discriminate the suboptimal arms.
Long phases $=$ the variance of the mean estimators of good arms is increasing with the length of the early phases

$$H_{\mathrm{P1}}(\boldsymbol{a}) \triangleq \max_{k \in [K]} \frac{\sum_{i=\langle k \rangle}^{K} (a_i - a_{i+1}) i + \frac{1}{24} K a_{\langle k \rangle} \Delta_k}{a_{\langle k \rangle}^2 \Delta_k^2} \overline{\log} K$$

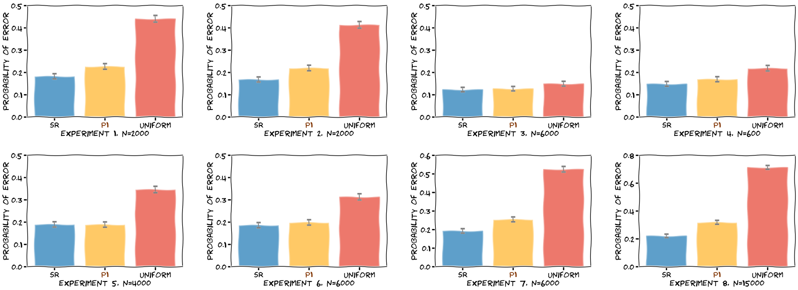
$$H_{\mathrm{P1}} \triangleq \min_{\boldsymbol{a} \in \boldsymbol{A}} H_{\mathrm{P1}}(\boldsymbol{a}).$$

Solution: Set $T_i = T \frac{\Delta_1}{\Delta_i}$ as in the lower bound.

Corollary The complexity H_{P1} of $\mathrm{P1}$ matches the complexity H_{BOB} from the lower bound of Theorem 4 of up to log factors,

$$H_{\mathrm{P1}} = \mathcal{O}\left(H_{\mathrm{BOB}} \overset{2}{\log} K \right).$$

Experimental setup	H_{SR}	H_{BOB}	H_{UNIF}
1. 1 group of bad arms	2000	2000	2000
2. 2 groups of bad arms	1389	2083	3125
3. Geometric progression	5540	5540	11080
4. 3 groups of bad arms	400	500	938
5. Arithmetic progression	3200	3200	24000
6. 2 good, many bad arms	5000	7692	50000
7. 3 groups of bad arms	4082	5714	12000
8. Square-root gaps	3200	22627	160000



Empirical behavior above mimics the behavior of the complexities in the table.

Thank you!

Bibliography

Robin Allesiardo, Raphael Féraud, and Odalric-Ambrym Maillard. The non-stationary stochastic multi-armed bandit problem. International Journal of Data Science and Analytics, 2017.

Jason Altschuler, Victor-Emmanuel Brunel, and Alan Malek. Best-arm identification for contaminated bandits. arXiv preprint arXiv:1802.09514, 2018.

Jean-Yves Audibert, Sebastien Bubeck, and Rémi Munos. Best-arm identification in multi-armed bandits. In Conference on Learning Theory, 2010.

Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In Conference on Learning Theory 2016

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. SIAM Journal on Computing, 32(1), 2002.

Sebastien Bubeck and Aleksandrs Slivkins. The best of both worlds: stochastic and adversarial bandits. In Conference on Learning Theory, 2012.

Alexandra Carpentier and Andrea Locatelli. Tight (lower) bounds for the fixed budget best-arm identification bandit problem. In Conference on Learning Theory, 2016.

Kevin Jamieson and Ameet Talwalkar. Non-stochastic best-arm identification and hyperparameter optimization. In International Conference on Artificial Intelligence and Statistics, 2016.

Oded Maron and Andrew Moore. Hoeffding Races: Accelerating model-selection search for classification and function approximation. In Neural Information Processing Systems, 1993.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In International Conference on Machine Learning, 2014.

Julian Zimmert, Yevgeny Seldin. Tsallis-INF: An Optimal Algorithm for Stochastic and Adversarial Bandits, JMLR 2021.