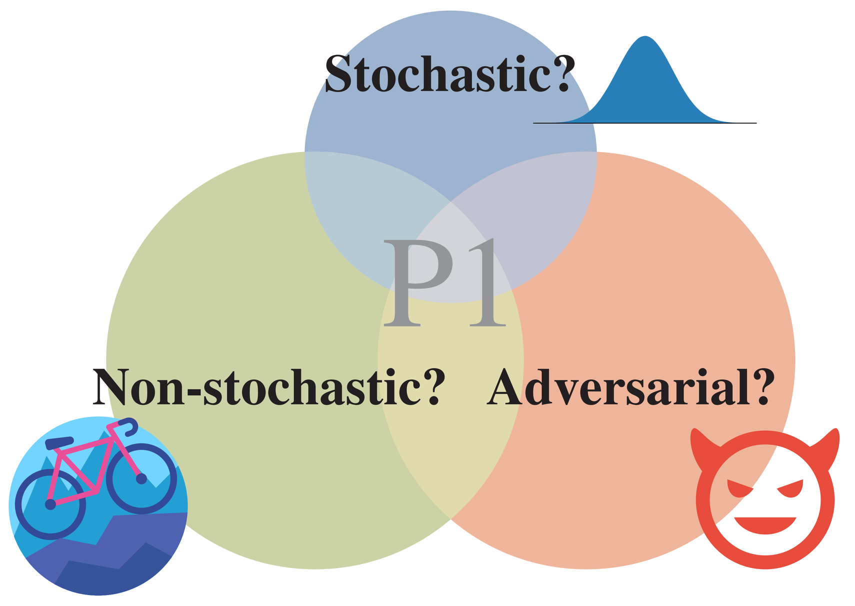
BEST OF BOTH WORLDS: STOCHASTIC & ADVERSARIAL BEST-ARM IDENTIFICATION

YASIN ABBASI-YADKORI, PETER BARTLETT, VICTOR GABILLON, ALAN MALEK & MICHAL VALKO

WHAT GIVES?



Find your best option when the data is potentially non-stochastic or adversarial!

THE GAME: LEARNER VS ADVERSARY

For $t = 1, 2, \ldots, n$,
- simultaneously, Learner picks arm $I_t \in [K]$,
- 😈/▲ picks gain $g_t \in [0,1]^K$.
- Then, the learner observes g_{t,I_t}.

Recommend arm J_n hoping $J_n = k^\star$.

😈 Adversarial 😈	▲ Stochastic ▲
arbitrary $g_{k,t}$	$g_{k,t}$ sampled i.i.d. from ν_k
$k_g^\star = \arg\max_{k\in[K]} G_k$, $G_k = \sum_{t=1}^n g_{k,t}$	$k^\star_{\mathrm{STO}} = \arg\max_{k\in[K]} \mu_k$
$e_{\mathrm{ADV}}(n) \triangleq \mathbb{P}(J_n \neq k_g^\star)$	$e_{\mathrm{STO}}(n) \triangleq \mathbb{P}(J_n \neq k^\star_{\mathrm{STO}})$

Gaps:
$$n\Delta_k^g \triangleq \begin{cases} G_{(1)} - G_k & \text{if } k \neq k_g^\star, \\ G_{(1)} - G_{(2)} & \text{if } k = k_g^\star. \end{cases}$$

Notions of complexity:
$$H_{\mathrm{SR}} \triangleq \max_{k\in[K]} \frac{k}{\Delta_{(k)}^2} \quad \text{and} \quad H_{\mathrm{UNIF}} \triangleq \frac{K}{\Delta_{(1)}^2}.$$

OPTIMAL UNIFORM LEARNER

`Rule`: I_t uniformly at random.

Th 1 (`Rule` vs. 😈). *For all n, adversarial g,*
$$e_{\mathrm{ADV}(g)}(n) = \mathcal{O}\left(\exp\left(-\frac{n}{H_{\mathrm{UNIF}(g)}}\right)\right).$$

Th 2 (😈 Lower bound). *For any learner, a g^1 of complexity H_{UNIF},*
$$e_{g^1}(n) = \Omega\left(\exp\left(-\frac{n}{H_{\mathrm{UNIF}}}\right)\right).$$

`Rule`: optimal gap-dependent rates against 😈.

¿BEST OF BOTH WORLDS? (BOB)

Existing robust solutions?

	$e_{\mathrm{STO}}(n)$		$e_{\mathrm{ADV}(g)}(n)$	
SR [1]	✔	$e^{\frac{-n}{H_{\mathrm{SR}} \log K}}$	✘	1
`Rule`	✘	$e^{\frac{-n}{H_{\mathrm{UNIF}}}}$	✔	$e^{\frac{-n}{H_{\mathrm{UNIF}}}}$

***BOB question:** A learner performing optimally in **both** the stochastic and adversarial cases while not being aware of the nature of the rewards ?*

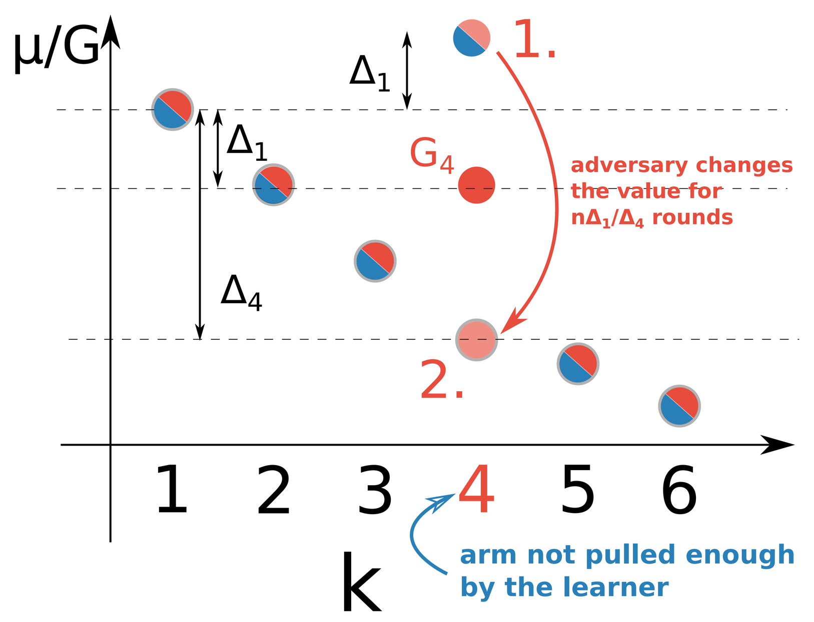
Why is the BOB question challenging?
- Bias of estimator $\widehat{G}_{k,t} = \frac{t\sum_{t'=1}^t \mathbb{1}\{I_{t'}=k\} g_{k,t'}}{\sum_{t'=1}^t \mathbb{1}\{I_{t'}=k\}}$
- Variance of $\widetilde{G}_{k,t} = \sum_{t'=1}^t \frac{g_{k,t'}}{p_{k,t'}} \mathbb{1}\{I_{t'} = k\}$

Pull uniformly for too long and incur a large variance of order K in $\widetilde{G}_{k,t}$.

¡IMPOSSIBLE BOB!

New notion of complexity
$$H_{\mathrm{BOB}} \triangleq \frac{1}{\Delta_{(1)}} \max_{k\in[K]} \frac{k}{\Delta_{(k)}}.$$

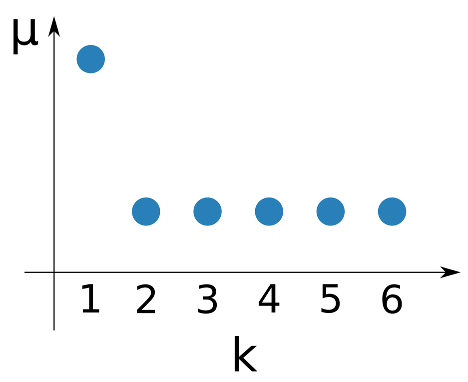
Th 3 (Lower bound for the BOB challenge). *For any learner, for any H_{BOB} there exists an stochastic problem with complexity H_{BOB} such that*
$$\textit{if} \quad e_{\mathrm{STO}}(n) \le \frac{1}{64}\exp\left(-\frac{2048n}{H_{\mathrm{BOB}}}\right),$$
***then** there exists an adversarial problem where*
$$e_{\mathrm{ADV}(g)}(n) \ge \frac{1}{16}.$$



DIFFERENT REGIMES

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

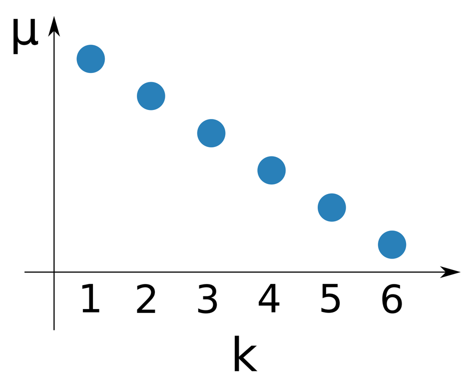
- **Flat regime**



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

BOB is achieved by `Rule`.

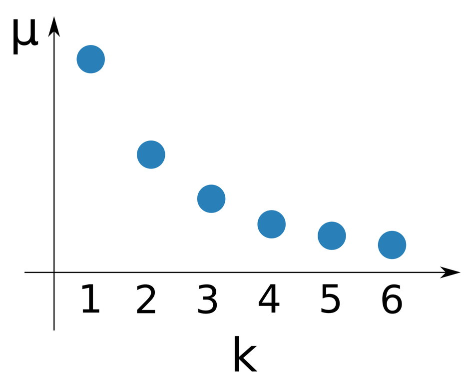
- **Linear regime**



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

BOB can be achieved but not by `Rule`.
We 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!

THE BYCWF (THE BEST YOU CAN WISH FOR)

Theorem 1 (Upper bounds for P1). *For any problems:*
- $$e_{\mathrm{STO}}(n) = \mathcal{O}\left(\exp\left(-\frac{n}{H_{BOB}\log^2(K)}\right)\right)$$
- $$e_{\mathrm{ADV}(g)}(n) = \mathcal{O}\left(\exp\left(-\frac{n}{\log(K) H_{\mathrm{UNIF}(g)}}\right)\right)$$

[1] J.-Y. Audibert, S. Bubeck, and R. Munos. Best-arm identification in multi-armed bandits. In *Conference on Learning Theory*, 2010.

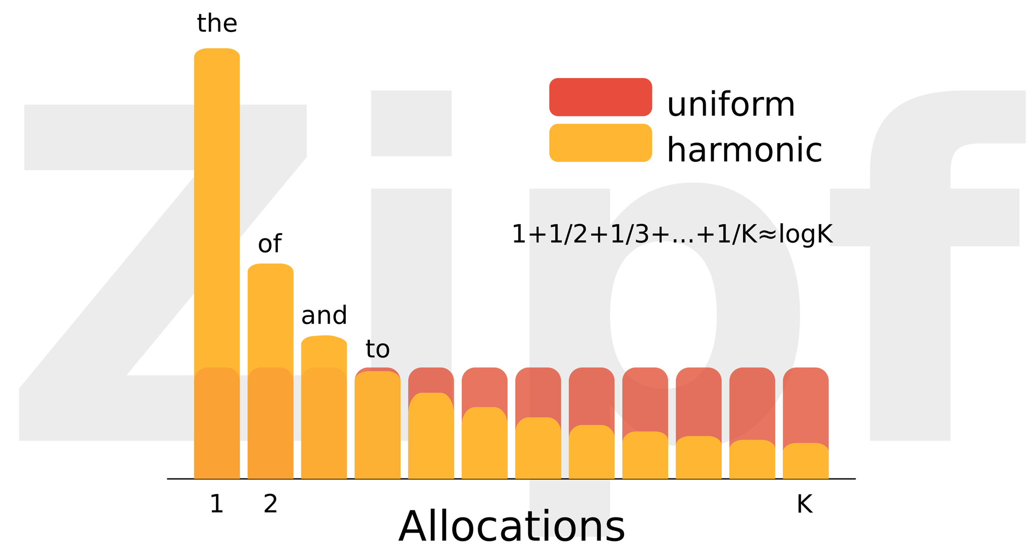
THE P1 ALGORITHM

`P1` pulls
- the $\widehat{best}$ arm with probability 1
- the second $\widehat{best}$ arm with proba $\frac{1}{2}$
- the third $\widehat{best}$ arm with probability $\frac{1}{3}$
- and so on ... (and normalize)

For $t = 1, 2, \ldots$
- Sort & rank arms by decreasing $\widetilde{G}_{\cdot,t-1}$: Rank arm k as $\widetilde{\langle k \rangle}_t \in [K]^a$.
- Select I_t with $\mathbb{P}(I_t = k) \triangleq \frac{1}{\widetilde{\langle k \rangle}_t \overline{\log} K}$.

Recommend, $J_t \triangleq \arg\max_{k\in[K]} \widetilde{G}_{k,t}$.

[a] Brake arbitrarily any problematic comparisons.

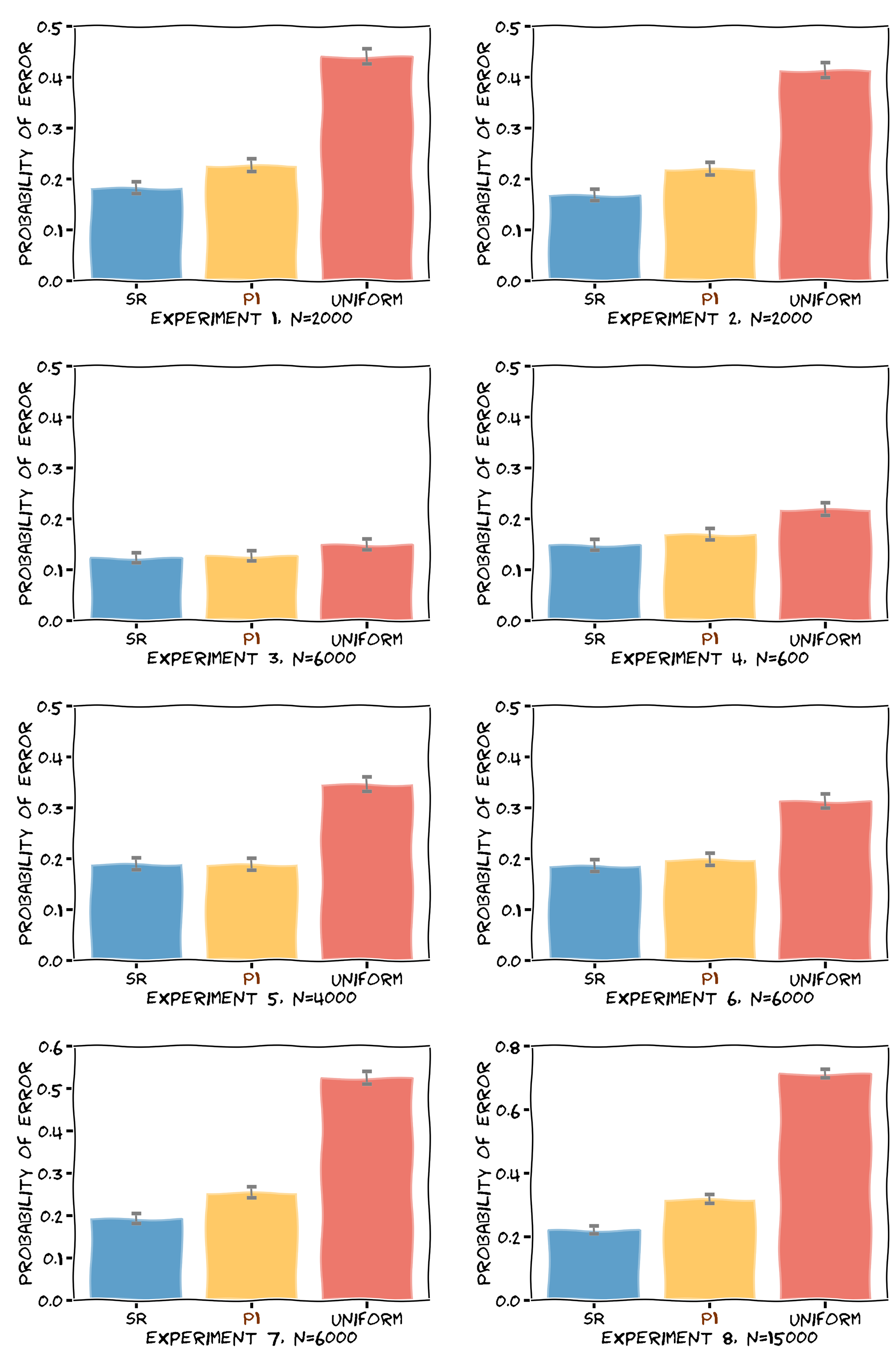


W.r.t. `Rule`**,** `P1` **early bets are almost costless!**

`P1` follows the allocation proportions of `SR`[1]

STOCHASTIC CASE EXPERIMENTS

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 prog	5540	5540	11080
4. 3 groups of bad arms	400	500	938
5. Arithmetic prog	3200	3200	24000
6. 2 good, many bad	5000	7692	50000
7. 3 groups of bad arms	4082	5714	12000
8. Square-root gaps	3200	22M	160M



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