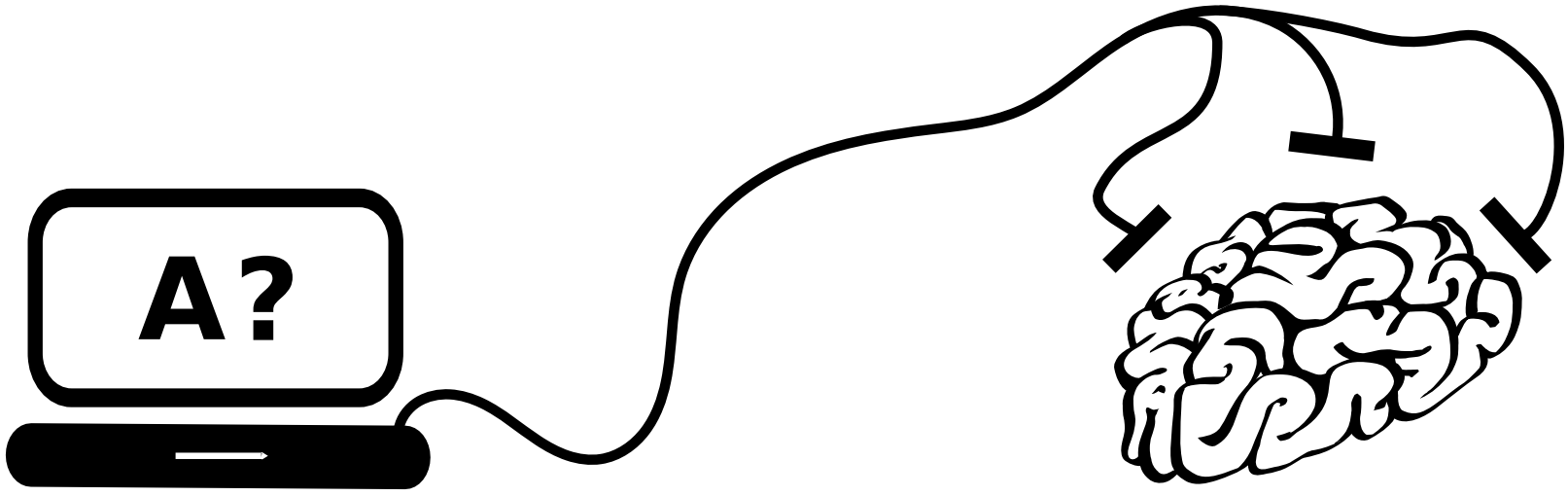
Multi-Bandit Best Arm Identification

V. Gabillon, M. Ghavamzadeh, A. Lazaric & S. Bubeck

Sequel Group Meeting, 21 octobre, 2011.

An Example

A Brain-Computer Interface (BCI) problem



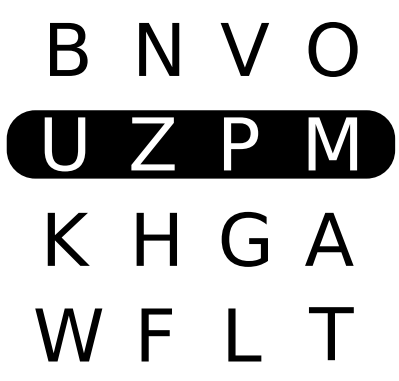
- A **computer** tries to guess the letter chosen by a **user**.
- It asks a **series** of **n** questions to the user.
- The answers are the recorded **noisy** brain activity signals.

Grid of letter shown to the user:

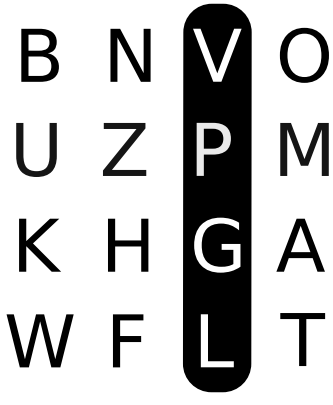
B N V O
U Z P M
K H G A
W F L T

Two types of questions:

1.) Is your letter in this **row**?



2.) Is your letter in this **column**?



- Which questions should be asked to best identify the letter?

A new bandit problem...

- Finding the right **row** is a multi-armed bandit problem.
- Finding the right **column** is a multi-armed bandit problem.

New: Solving **several** bandit problems at the same time.

The loss:

$$\begin{cases} 1 & \text{if } \textbf{all} \text{ the bandits are solved,} \\ 0 & \text{otherwise.} \end{cases}$$

... based on pure exploration problems.

Known algorithms in the single bandit case.

Outline

Best arm identification

Identifying the best arm:

- **K** arms: $\nu_1, \nu_2, \ldots \nu_K$ distributions (arms) in $[0, b]$.
- At each time step t the player chooses an arm $I(t)$ to pull and receives a sample from $\nu_{I(t)}$
- At time **n** the player has to recommend an arm **J(n)**.
- The player wins if he has recommended the distribution with highest mean k^*.

What is the strategy of allocation which minimizes the error?

$$\ell(n) = P(error) = P(J(n) \neq k^*)$$

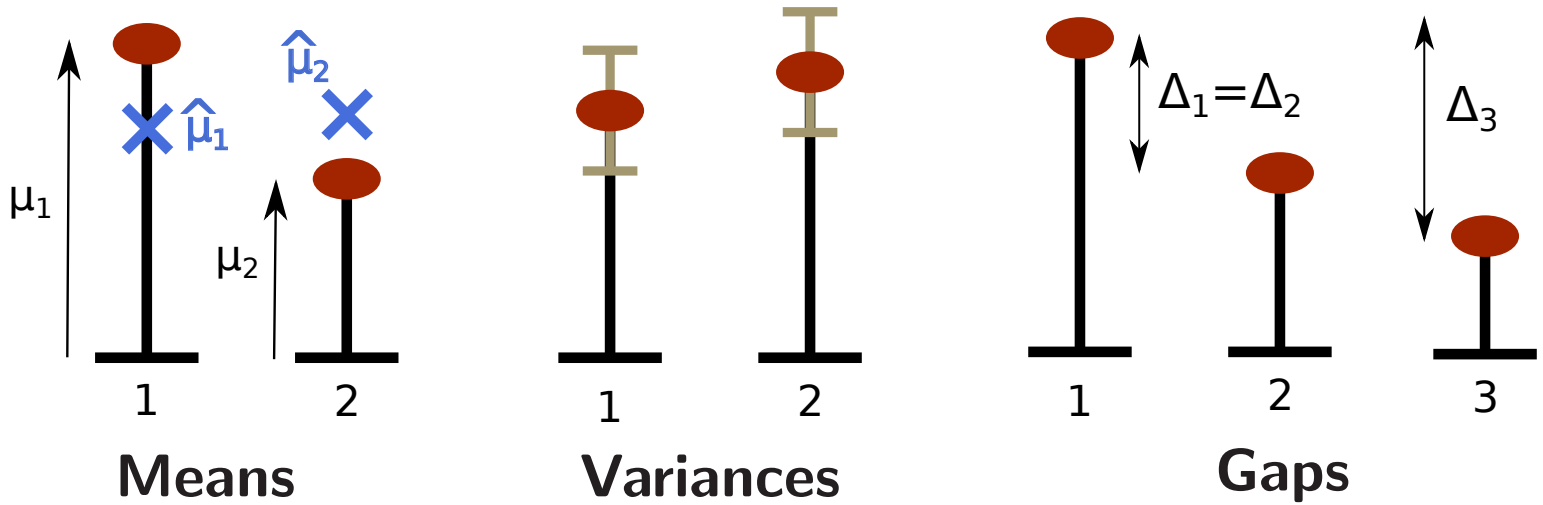
Notations

In arm k:

- **Number of pulls**: $\mathbf{T_k(t)}$ (at time t).
- **Mean-variance**: $\mu_\mathbf{k}$ and $\sigma^2_\mathbf{k}$.
- **Estimated mean & variance**:

$$\widehat{\mu}_\mathbf{k}(t) = \frac{1}{T_k(t)} \sum_{s=1}^{T_k(t)} X_{k,s} \quad \text{and} \quad \widehat{\sigma^2}_\mathbf{k}\mathbf{(t)} = \frac{1}{T_k(t)-1} \sum_{s=1}^{T_k(t)} (X_{k,s} - \widehat{\mu}_k(t))^2$$

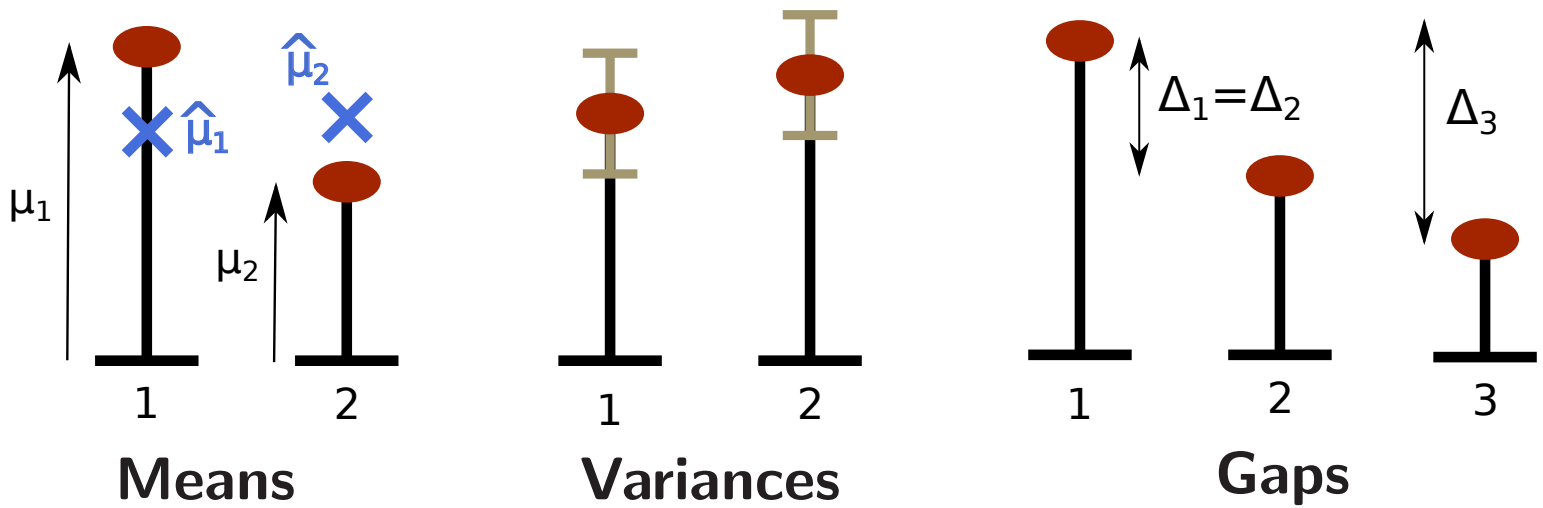
$X_{k,s}$ is the s^{th} sample from the arm.



Notations

In arm k:

- **Best arm**: $\mathbf{k^*} = \arg\max_k \mu_k$.
- **Gap**: $\mathbf{\Delta_k} = |\max_{j \neq k} \mu_j - \mu_k|$.



Simple Regret vs. Cumulative regret

Cumulative regret

The goal is to minimize:

$$R(n) = n\mu_{k^*} - \sum_{t=1}^{n} X_{I(t)}$$

Dilemma: Spot the best arm (exploration) & exploit it (exploitation).

Simple regret

$$r(n) = \mu_k^* - \mu_{J(n)}$$

Pure exploration: No more exploration vs. exploitation trade-off.

$$\mathbb{E}r(n) \leq b \times \ell(n),$$

Budgeted Learning

The total number of pulls n is fixed and known by the player.

Here: we study how to optimize the quality of the recommendation with a fixed budget

vs.

Trying to optimize the budget in order to achieve a given confidence level (Hoeffding Races, **Even-Dar et.al. (2002)** ...).

Uniform allocation

 # Divide the budget equally over the arm:
Pull every arm n/K times.

 # Output the arm with highest estimated mean:
Output $J(n) \in \arg\max_k \widehat{\mu}_k(n)$.

What is the **probability of error** of the uniform allocation?

Distinguishing two arms pulled equally:

with $\mu_1 > \mu_2$ and $\Delta = \mu_1 - \mu_2$

$$\ell(n) := \mathbb{P}\,(error) = \mathbb{P}\,(\,\widehat{\mu}_1(n) \;<\; \widehat{\mu}_2(n)\,) \leq exp\left(\; -\frac{n}{2}(\frac{\Delta}{b})^2\;\right)$$

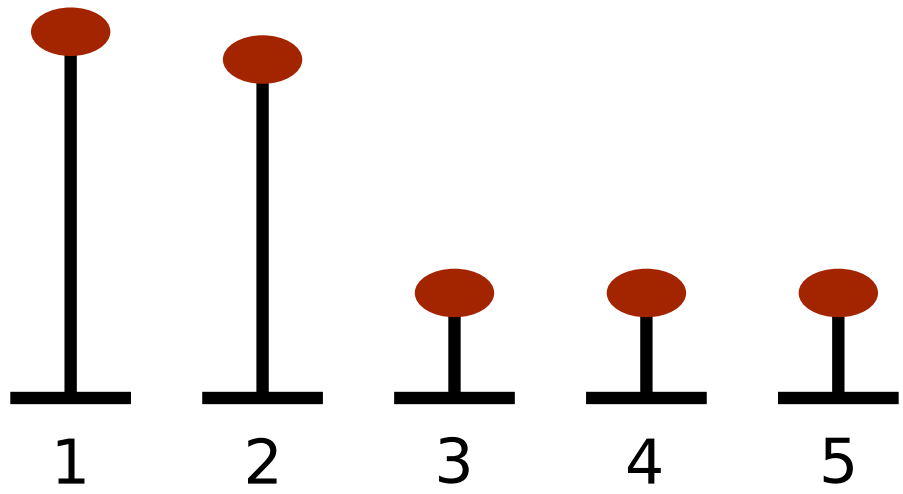
For a fixed accuracy, the arms have to be pulled of order $\left(\frac{\mathbf{b}}{\mathbf{\Delta}}\right)^{\mathbf{2}}$.

Error of Uniform allocation:

$$\ell_{Unif}(n) := \mathbb{P}\,(\,J(n) \neq k^*\,) \leq \sum_{k \neq k^*} exp\left(\; -\frac{n}{K}(\frac{\Delta_k}{b})^2\;\right)$$

$$\leq K\; exp\left(\; -\frac{n}{K}(\frac{\min_k \Delta_k}{b})^2\;\right)$$

Smarter than uniform allocation?



- We should not **waste** to much pulls on arms 3,4,5.
- We shoud focus on 1 and 2 to discriminate them.

UCB-Exploration (UCB-E)

Audibert, Bubeck & Munos (2010)

Parameter: exploration parameter a, range b, number of pulls n.

Index of arm k**:**

$$B_k(t) = \widehat{\mu}_k(t-1) + b\sqrt{\frac{a}{T_k(t-1)}}$$

for each round $t = 1, 2, \ldots, n$ **do**
  # Explore the action with the highest index:
  Draw $I(t) \in \arg\max_{k \in \{1,\ldots,K\}} B_k(t)$
end for

 # Output the arm with highest estimated mean:
Output $J(n) \in \arg\max_k \widehat{\mu}_k(n)$.

UCB-Exploration (UCB-E)

Audibert, Bubeck & Munos (2010)

Theorem: **Bound on the probability of error**

If $a = \frac{n}{H}$,

$$\ell_{UCB-E}(n) \leq 2\, n\, K\, exp\left(- O(\frac{n}{\mathbf{H}})\right)$$

and $T_k(n) \approx O(\frac{b}{\Delta_k^2}) \quad \forall k$.

UCB-E pulls every arms according to its complexity.

Complexity H:

$$H = \sum_k (\frac{b}{\Delta_k})^2$$

The global complexity is the **sum** of the complexity of the arms.

Comparing Uniform vs UCB-E

Probability of errors:

$$\ell_{Uniform}(n) \quad \text{vs} \quad \ell_{UCB-E}(n)$$

$$exp\left(-\frac{n}{K}\left(\frac{\min_k \Delta_k}{b}\right)^2 \right) \qquad exp\left(-\frac{n}{H} \right) \quad \text{\# Comparing the bounds}$$

$$K \max_k \frac{1}{\Delta_k^2} \geq \sum_k \frac{1}{\Delta_k^2}$$

- UCB-E is expected to improve upon Uniform when the gaps are differents.
- This was supported by numerical simulations.

Comparing Uniform vs UCB-E

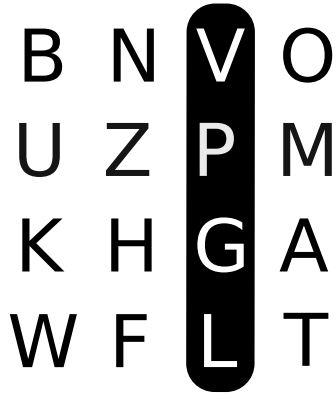
Probability of errors:

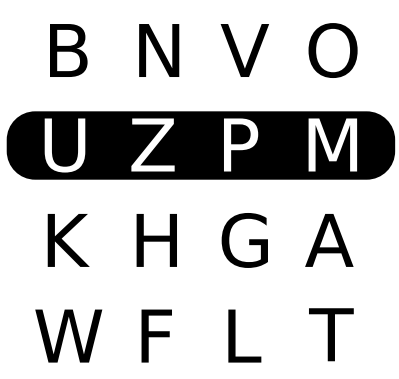
$$\ell_{Uniform}(n) \quad \text{vs} \quad \ell_{UCB-E}(n)$$

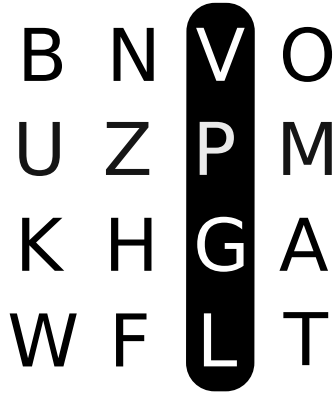
$$exp\left(-\frac{n}{K}\left(\frac{\min_k \Delta_k}{b}\right)^2\right) \leq exp\left(-\frac{n}{H}\right) \quad \text{\# Comparing the bounds}$$

$$K \max_k \frac{1}{\Delta_k^2} \geq \sum_k \frac{1}{\Delta_k^2}$$

- UCB-E is expected to improve upon Uniform when the gaps are differents.
- This was supported by numerical simulations.







New Goal: To solve **several** problems at the same time.

Outline

Multi-bandit pure exploration problem

- **M** bandits
- **K** arms in each bandit.
- At each time step t the player chooses a **bandit-arm pair** $I(t)$ to pull and receives a sample from $\nu_{I(t)}$
- At time n the player **recommends an arm for each bandit**.
- The player wins if he has recommended the distribution with highest mean **in all the bandits**.

What is the strategy of allocation which minimizes the error?

Issues

- Is UCB-E a solution for this problem?
- What type of algotithm for bandit problem : A two stage algorithm?
  Allocation over bandits + allocation over arms.
- What is the complexity of this problem?

New notations

In bandit m and arm k:

- **Number of pulls**: $\mathbf{T_{mk}(t)}$ (at time t).
- **Mean-variance**: $\boldsymbol{\mu_{mk}}$ and $\boldsymbol{\sigma^2_{mk}}$.
- **Estimated mean & variance**:

$$\widehat{\boldsymbol{\mu}}_{\mathbf{mk}}(t) = \frac{1}{T_{mk}(t)} \sum_{s=1}^{T_{mk}(t)} X_{mk,s}$$

$$\widehat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\mathbf{mk}}\mathbf{(t)} = \frac{1}{T_{mk}(t)-1} \sum_{s=1}^{T_{mk}(t)} (X_{mk,s} - \widehat{\mu}_{mk}(t))^2$$

  $X_{mk,s}$ is the s^{th} sample from arm.

- **Best arms**: $\mathbf{k^*_m} = \arg\max_k \mu_{mk}$ and $\widehat{\mathbf{k}}^{\mathbf{*}}_{\mathbf{m}} = \arg\max_k \widehat{\mu}_{mk}$.
- **Gap** of arm k: $\boldsymbol{\Delta_{mk}} = |\max_{j \neq k} \mu_{mj} - \mu_{mk}|$.

UCB-E is not a good solution

On this problem,

- UCB-E will focus on the arms of bandits 1.
- UCB-E will **fail** to solve the two bandit problems.

Uniform strategies

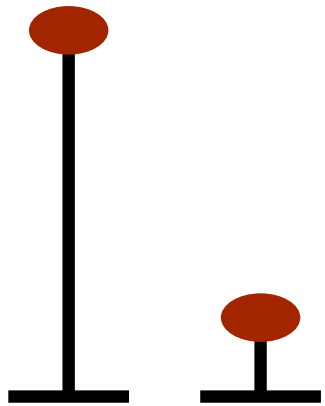
- **Uniform:** Pulls every arm n/MK.
- **Uniform + UCB-E:**
  1. The number of pulls for each bandit is n/M.
  2. Play UCB-E in each bandit.

Bounds on the errors: $\ell(n) := \mathbb{P}\left(\exists m : J_m(n) \neq k_m^*\right)$

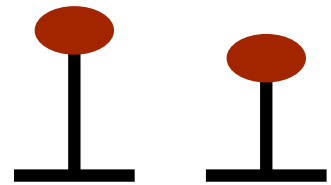
$$\ell_{\text{Uniform}}(n) \leq \exp\left(-O\left(\frac{n}{MK \max_{m,k} H_{mk}}\right)\right), \quad H_{mk} = \left(\frac{b}{\Delta_{mk}}\right)^2,$$

$$\ell_{\text{Uniform+UCBE}}(n) \leq \exp\left(-O\left(\frac{n}{M \max_{m} H_{m}}\right)\right), \qquad H_m = \sum_k H_{mk}.$$

Allocation over bandits



bandit 1 : **easy**

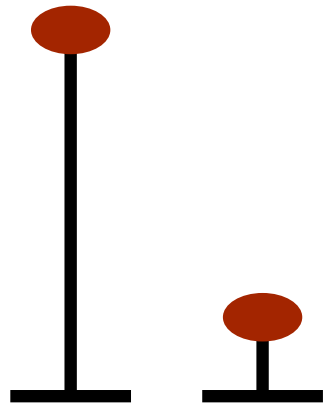


bandit 2: **hard**

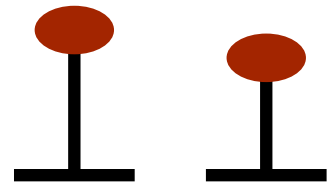
On this problem,

- More pulls should be allocated on bandit 2.
- Uniform allocations are not satisfactory.

Allocation other states



bandit 1 : **easy**



bandit 2: **hard**

Allocating according to the gaps?

Allocation other states

bandit 1 : **easy** bandit 2: **hard**

Gaps:

- First idea: $\widehat{\Delta}_{mk} = \max_j \widehat{\mu}_{mj} - \widehat{\mu}_{mk}$ **BUT:** $\widehat{\Delta}_{m\widehat{k}^*_m} = 0$
- Second idea: $\widehat{\Delta}_{mk} = |\max_{j \neq k} \widehat{\mu}_{mj} - \widehat{\mu}_{mk}|$

From UCB-E to GapE

Parameter: exploration parameter a, range b, number of pulls n.

Index of arm k**:**

$$B_k(t) = \widehat{\mu}_k(t-1) + b\sqrt{\frac{a}{T_k(t-1)}}$$

for each round $t = 1, 2, \ldots, n$ **do**
  # Explore the action with the highest index:
  Draw $I(t) \in \arg\max_{k \in \{1,\ldots,K\}} B_k(t)$
end for

 # Output the arm with highest estimated mean:
Output $J(n) \in \arg\max_k \widehat{\mu}_k(n)$.

From UCB-E to GapE

Parameter: exploration parameter a, range b, number of pulls n.

Index of arm k:

$$B_k(t) = \widehat{\mu}_k(t-1) - \max_k \widehat{\mu}_k(t-1) + b\sqrt{\frac{a}{T_k(t-1)}}$$

for each round $t = 1, 2, \ldots, n$ **do**
  # Explore the action with the highest index:
  Draw $I(t) \in \arg\max_{k \in \{1,\ldots,K\}} B_k(t)$
end for

 # Output the arm with highest estimated mean:
Output $J(n) \in \arg\max_k \widehat{\mu}_k(n)$.

From UCB-E to GapE

Parameter: exploration parameter a, range b, number of pulls n.

Index of arm k**:**

$$B_k(t) = -\widehat{\Delta}_k(t-1) + b\sqrt{\frac{a}{T_k(t-1)}}$$

for each round $t = 1, 2, \ldots, n$ **do**
  # Explore the action with the highest index:
  Draw $I(t) \in \arg\max_{k \in \{1,\ldots,K\}} B_k(t)$
end for

 # Output the arm with highest estimated mean:
Output $J(n) \in \arg\max_k \widehat{\mu}_k(n)$.

From UCB-E to GapE

Parameter: exploration parameter a, range b, number of pulls n.

Index of arm k:

$$B_k(t) = -\widehat{\Delta}_{mk}(t-1) + b\sqrt{\frac{a}{T_{mk}(t-1)}}$$

for each round $t = 1, 2, \ldots, n$ **do**

 # Explore the action with the highest index:

Draw $I(t) \in \arg\max_{k \in \{1,\ldots,K\} \& m \in \{1,\ldots,M\}} B_{mk}(t)$

end for

 # Output the arms with highest estimated mean:

Output $J_m(n) \in \arg\max_k \widehat{\mu}_{mk}(n)$.

Gap-based exploration (GapE)

Theorem: **Bound on the probability of error**

If $a = \frac{n}{H}$,

$$\ell_{GapE}(n) \leq 2\, n\, M\, K\, exp\left(- O(\frac{n}{\mathbf{H}}) \right)$$

and $T_{mk}(n) \approx O(\frac{b}{\Delta_{mk}^2}) \quad \forall k, \forall m.$

GapE pulls every arms according to its complexity.

Complexity H:

$$H = \sum_m \sum_k (\frac{b}{\Delta_{mk}})^2$$

The global complexity is the **sum** of the complexity of the arms.

Comparing Uniform allocations vs GapE

Probability of errors:

$$\ell_{Uniform}(n) \quad \text{vs} \quad \ell_{Unif+UCB-E}(n) \quad \text{vs} \quad \ell_{GapE}(n)$$

$$\exp\left(-\frac{n}{MK \max_{m,k} H_{mk}}\right) \qquad \exp\left(-\frac{n}{M \max_{m} H_{m}}\right) \qquad exp\left(-\frac{n}{H}\right)$$

$$MK \max_{m,k} \frac{1}{\Delta_{mk}^2} \quad \geq \quad M \max_{m} \sum_{k} \frac{1}{\Delta_{mk}^2} \quad \geq \quad \sum_{m} \sum_{k} \frac{1}{\Delta_{mk}^2}$$

- GapE is expected to improve upon Uniform allocations when the gaps are differents.
- This will be supported by numerical simulations.

Comparing Uniform allocations vs GapE

Probability of errors:

$$\ell_{Uniform}(n) \quad \text{vs} \quad \ell_{Unif+UCB-E}(n) \quad \text{vs} \quad \ell_{GapE}(n)$$

$$\exp\left(-\frac{n}{MK \max_{m,k} H_{mk}}\right) \leq \exp\left(-\frac{n}{M \max_{m} H_{m}}\right) \leq exp\left(-\frac{n}{H}\right)$$

$$MK \max_{m,k} \frac{1}{\Delta_{mk}^2} \geq M \max_{m} \sum_{k} \frac{1}{\Delta_{mk}^2} \geq \sum_{m} \sum_{k} \frac{1}{\Delta_{mk}^2}$$

- GapE is expected to improve upon Uniform allocations when the gaps are differents.
- This will be supported by numerical simulations.

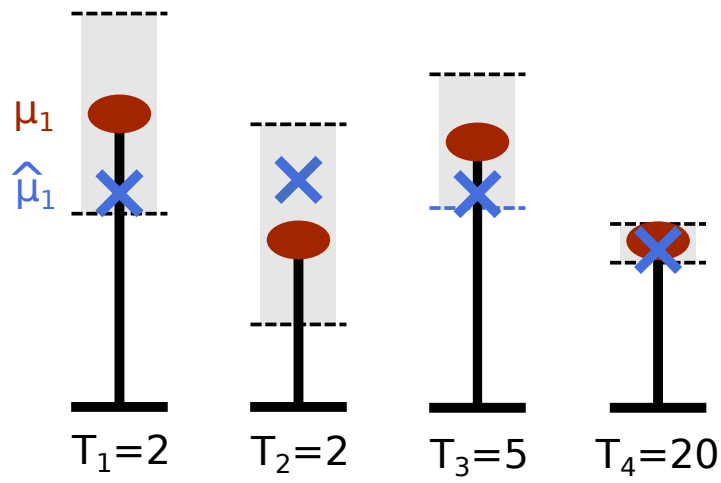
Sketch of the Proof:

Difficulty: Having upper bound on the estimated mean of the estimated best arm.

1. Characterize an event on which the algorithm succeeds and compute its probability.
2. On this event, give a condition on the number of pulls so that the algorithm succeeds.
3. Prove, on that event, an induction formula for each time t showing the dependence between the pulls of each arm and their gaps.
4. With the induction formula, compute the number of pulls at time n and show that it satisfies the requirement to succeed.

1.) The event where GapE succeeds

$$\mathcal{E} = \Bigg\{ \forall m \in \{1,\dots,M\},\ \forall k \in \{1,\dots,K\},\ \forall t \in \{1,\dots,n\},$$
$$\left|\widehat{\mu}_{mk}(t) - \mu_{mk}\right| < bc\sqrt{\frac{a}{T_{mk}(t)}} \Bigg\}.$$



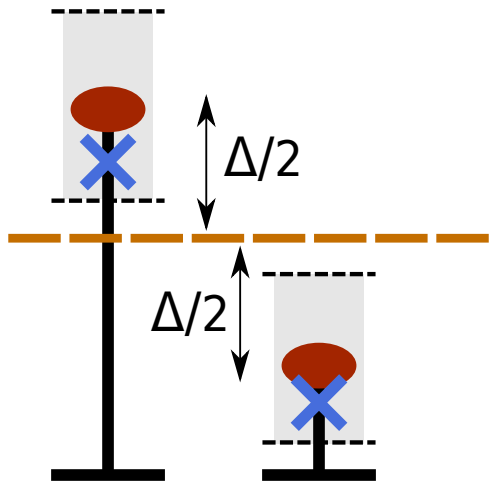
From Chernoff-Hoeffding's inequality and a union bound,

we have $\mathbb{P}(\xi) \geq 1 - 2MKn \exp\left(-2ac^2\right)$.

2.) Condition on the number of pulls

Success in arm k of bandit m if,

$$
\begin{aligned}
J_m(n) = k_m^* \quad &\Longleftrightarrow \quad \forall k, \quad \widehat{\mu}_{mk}(n) \leq \widehat{\mu}_{mk_m^*}(n) \\
&\Leftarrow \quad \forall k, \quad bc\sqrt{a/T_{mk}(n)} \leq \Delta_{mk}/2 \\
&\Leftrightarrow \quad \forall k, \quad T_{mk}(n) \geq \frac{4ab^2c^2}{\Delta_{mk}^2}
\end{aligned}
$$



3.) Induction formula

For any bandits (m, q) and arms (k, j), and for any $t \geq MK$,

$$-\Delta_{mk}+(1+d)b\sqrt{\frac{a}{\max\left(T_{mk}(t)-1,1\right)}} \geq -\Delta_{qj}+(1-d)b\sqrt{\frac{a}{T_{qj}(t)}},$$

It shows the dependence between the number of pulls of **all** the pair of arms from **all** the bandit.

4.) At time n

Proof by contradiction,

- Let us assume $\exists(m,k) : T_{mk}(n) \leq \frac{ab^2(1-d)^2}{\Delta_{mk}^2}$

$\Rightarrow -\Delta_{mk} + (1-d)b\sqrt{\frac{a}{T_{mk}(n)}} > 0$

$\Rightarrow -\Delta_{qj} + (1+d)b\sqrt{\frac{a}{T_{qj}(n)-1}} > 0, \ \forall(q,j)$ **induction formula**

$\Rightarrow T_{qj}(n) < \frac{ab^2(1+d)^2}{\Delta_{qj}^2} + 1$

- And we have

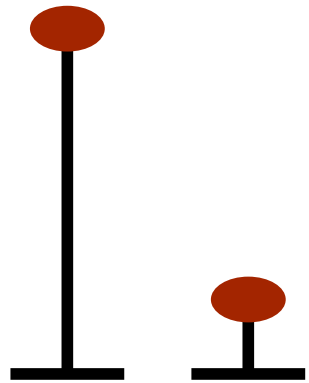
$$\sum_{q,j} T_{qj}(n) = n \quad \Rightarrow \quad n - MK < ab^2(1+d)^2 \sum_{q,j} \frac{1}{\Delta_{qj}^2}$$

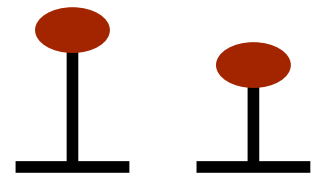
.

- So, if we select a such that $n - MK \geq ab^2(1+d)^2 \sum_{q,j} \frac{1}{\Delta_{qj}^2}$, we contradict the first assumption and this concludes the proof.

Extensions

The use of variance

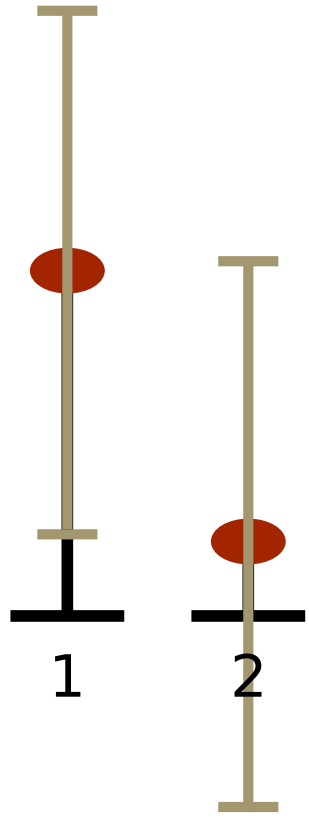




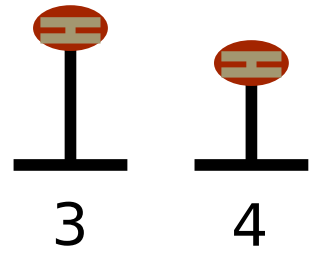
bandit 1: **easy**

bandit 2: **hard**

The use of variance

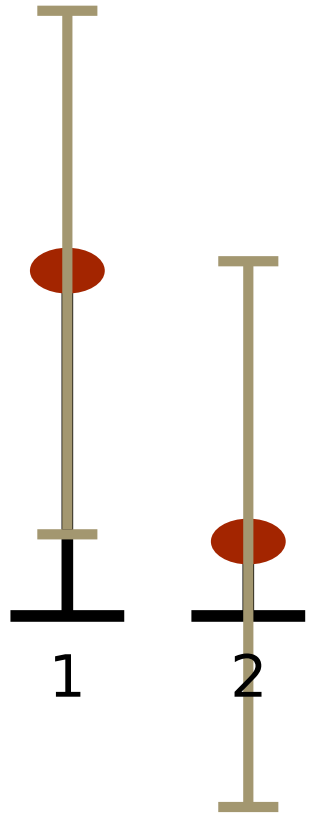


bandit 1: **?**

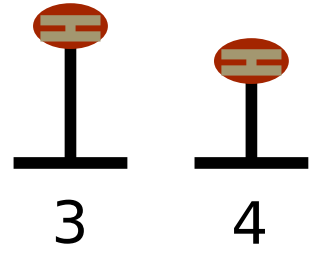


bandit 2: **?**

The use of variance



bandit 1: **hard**



bandit 2: **easy**

GapE-Variance (GapE-V)

Index of arm k:

$$B_k(t) = -\widehat{\Delta}_{mk}(t-1) + \sqrt{\frac{2a\,\widehat{\sigma}^2_{mk}(t-1)}{T_{mk}(t-1)}} + \frac{7ab}{3\Big(T_{mk}(t-1)-1\Big)}$$

Theorem: **Bound on the probability of error**

$$\text{If } a = \tfrac{n}{H^\sigma}, \qquad \ell_{GapE-V}(n) \leq 6\,n\,M\,K\,exp\left(-O(\tfrac{n}{H^\sigma})\right)$$

Variance-complexity H^σ:

$$H^\sigma = \sum_m \sum_k \frac{\left(\sigma_{mk} + \sqrt{\sigma^2_{mk} + (16/3)b\Delta_{mk}}\right)^2}{\Delta^2_{mk}}.$$

H^σ is smaller than H when the variance is small compared to b.

Adaptive version of GapE algorithms

Issues

- a should be tuned according to the complexities H and H^σ
- H and H^σ are rarely known in advance.

Online estimation of the complexities:

$$\mathrm{UCB}_{\Delta_i}(t) = \widehat{\Delta}_i(t-1) + \sqrt{\frac{1}{2T_i(t-1)}}$$

$$\widehat{H}(t) = \sum_{m,k} \frac{b^2}{\mathrm{UCB}_{\Delta_i}(t)^2},$$

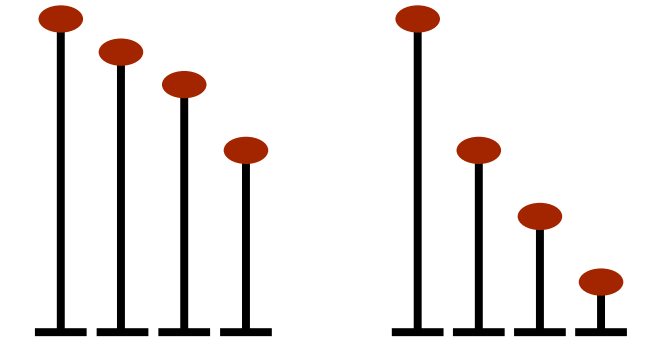
Outline

Parameters

- Theoretically, a should be proportional to $\frac{n}{H}$.
- In the experiments $a = \eta \frac{n}{H}$, η is a parameter to tune.
- η is suppose to correct the inaccuracy of the constants in the analysis
- the range of its nearly-optimal values should be constant across different problems

Comparison: GapE vs uniform strategies

Problem 1: $n = 700$



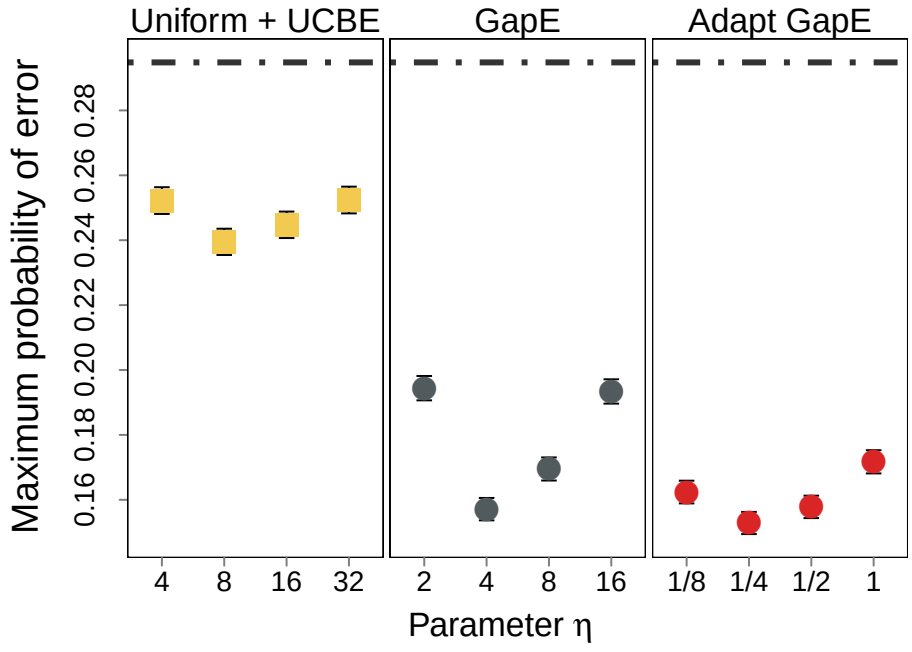
bandit 1: **H ≈ 900**

bandit 2: **H ≈ 70**

Allocation in percent in GapE:

[81]	[19]
[37, 36, 20, 7]	[36, 34, 17, 10]

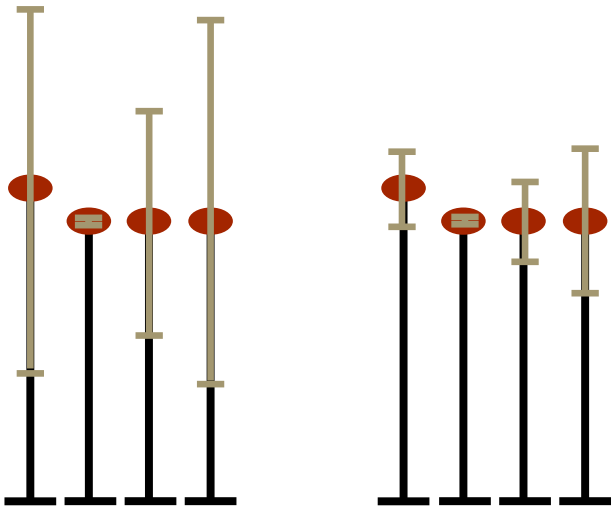
Results



- GapE allocates according to the complexities.
- GapE improves upon Uniform and Uniform+UCB-E.

Comparison: GapE vs GapE-V

Problem 2: $n = 1000$



bandit 1: **H ≈ 1600**, **H^σ ≈ 1400**

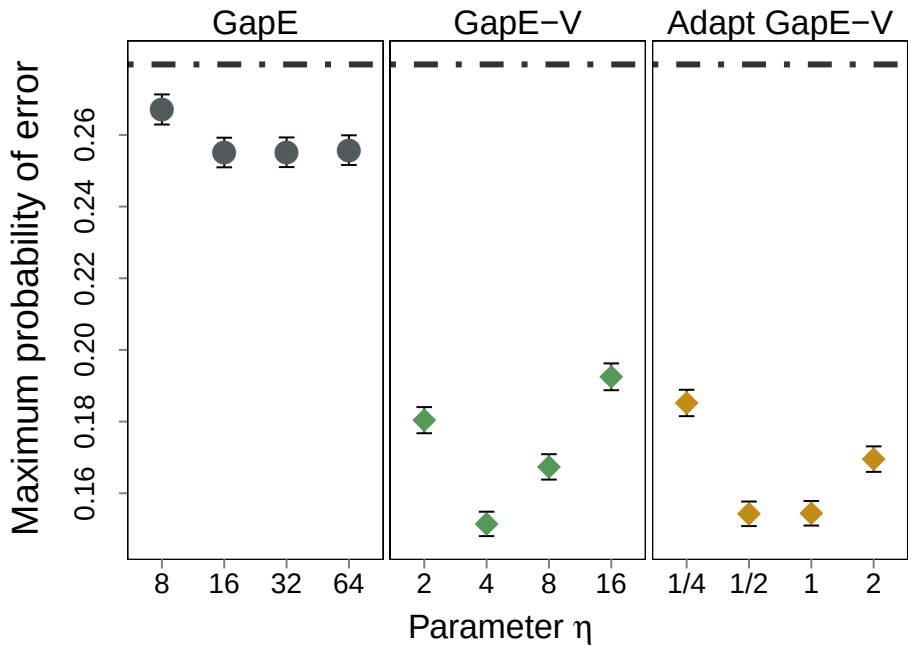
bandit 2: **H ≈ 1600**, **H^σ ≈ 600**

Allocation in % in GapE-V:

[79] [21]

[45, 8, 16, 29] [28, 9, 22, 40]

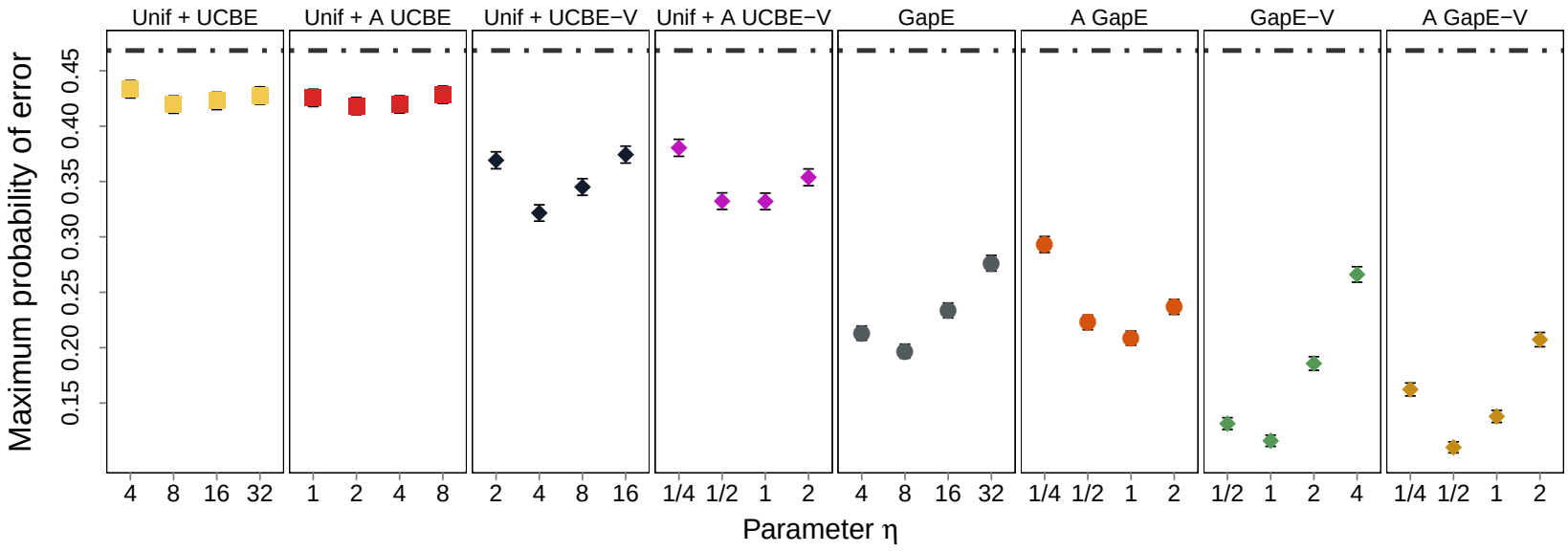
Results



- GapE-V allocates according to the variance-complexities.
- GapE-V improves upon GapE.

Final comparison

Problem 3 with strongly heterogeneous bandits & arms $M = K = 4$.



- 1^{st}: GapE-V, 2^{nd}: GapE, 3^{nd}: Unif+UCBEV, 4^{th}: Unif+UCBE
- Across the problems, in GapE-V, $\eta \in [1 - 4]$ is a near optimal choice.
- The **adaptive** versions of the algorithms have almost the same performance than their **"oracle"** counterpart.

Conclusions

- **New problem:** of best arm identification in a multi-bandit multi-armed setting.
- **New algorithm:** GapE, with an upper-bound for its probability of error showing its *potential* advantage over uniform strategies.
- **Extensions:**
  - Considering the variance
  - Adaptive version which estimates the complexity of the problem online.
- The **numerical simulations** show that GapE and GapE-V and their adaptive counterparts outperform other allocation strategies.

Perspectives

Application in RL?

In rollout allocation for classification-based policy iteration the goal is to identify the greedy action (*arm*) in each of the states (*bandit*) in a training set.

Dynamic vs Static strategy.

Gap-E is empirically better than a static strategy which would allocated according to the complexity. But they have the same theoretical guaranties. Can we improve something?

Other problems...

GapE-KL, data-dependent complexity, lower bound, pure exploration in structured set, simple regret, sum loss vs max loss, analysis of adaptive version...

Thank you!