

Best arm identification in fixed confidence MABs

Confidence intervals and asymptotic optimality

Jayakrishnan Nair

Department of Electrical Engineering, IIT Bombay

Multi-armed bandit problem

Fundamental problem in online learning: Learn the best among a basket of options (*a.k.a.*, *arms*) via sequential sampling



F_1



F_2



F_3



F_4

← unknown reward distributions


Example: Learn option (arm) with highest mean reward

Two flavours of Best Arm Identification (BAI) problem

Fixed budget setting

- Agent/algorithm has fixed budget of n samples/pulls
- After seeing n samples, algorithm outputs estimated best arm \hat{a}
- ***Goal: Design algorithms with the minimal probability of error, i.e., $P(\hat{a} \neq \text{best arm})$***

Fixed confidence setting (this talk)

- After each sample, algorithm must choose 
 - continue sampling**
 - stop**
- If algorithm stops, say at random stopping time τ algorithm and outputs estimated best arm \hat{a} , we require

$$P(\tau < \infty, \hat{a} \neq 1) \leq \delta$$

 *prescribed error threshold*

Algorithms satisfying this requirement are called ***sound/ δ -PC***

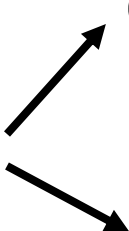
Goal: Design sound algorithms with the minimal $E[\tau]$

Fixed confidence setting (this talk)

Algorithm in this setting has three components:

- Stopping rule
- Sampling rule
- Recommendation rule

Broadly, two classes of algorithms



```
graph LR; A[Broadly, two classes of algorithms] --> B[Confidence Interval based]; A --> C[Track & Stop style]
```

Confidence Interval based

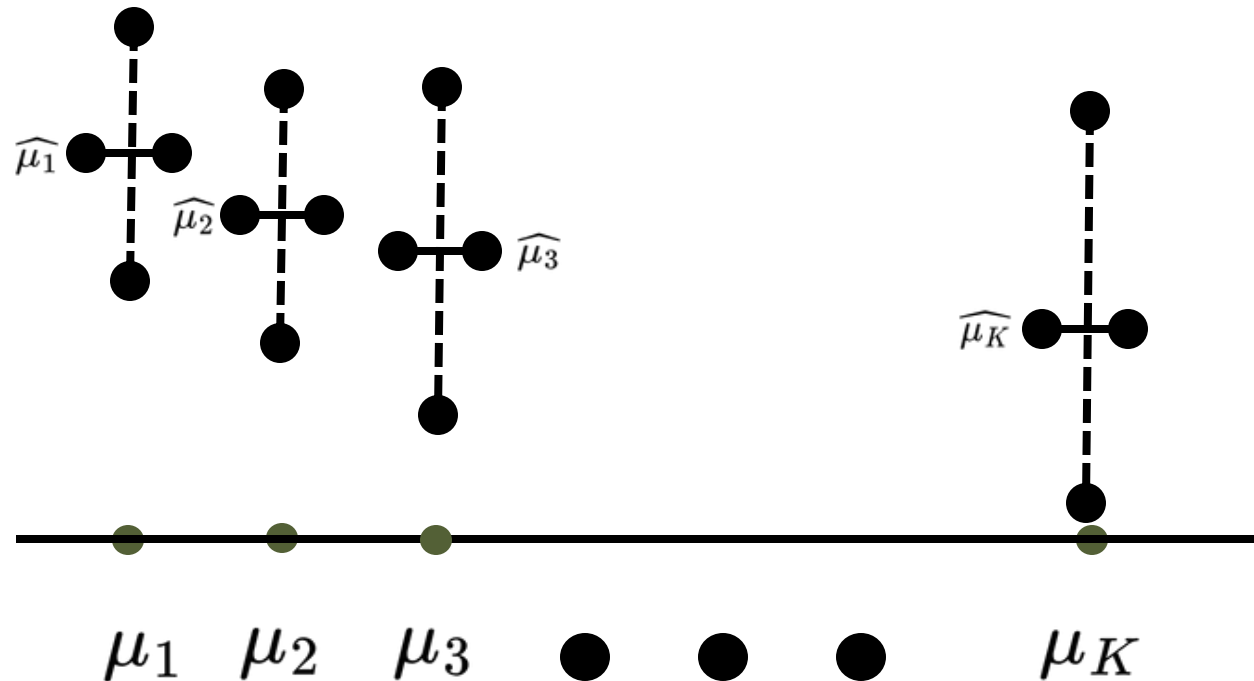
Track & Stop style

Confidence Interval based algorithms

- K arms
- Arm i has reward distribution ν_i (1-subGaussian), mean reward μ_i
- Bandit instance is $\nu = (\nu_i, 1 \leq i \leq K)$
- Assume $\mu_1 > \mu_2 \geq \mu_3 \geq \cdots \mu_K$

Confidence Interval based algorithms

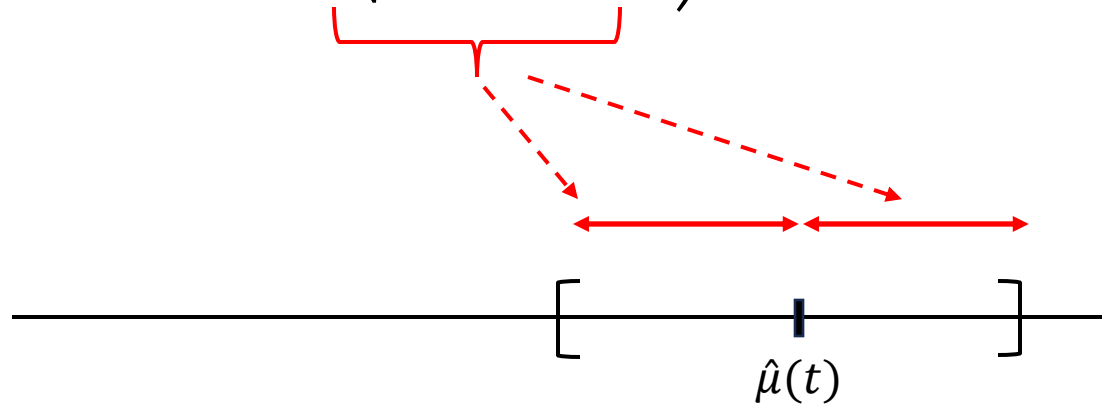
- Maintain (algo computable) confidence intervals on mean of each arm
- Use these to guide both sampling as well as stopping



SubGaussian concentration inequality:

$$P(|\hat{\mu}(t) - \mu| > \epsilon) \leq 2 \exp\left(-\frac{t\epsilon^2}{2}\right)$$

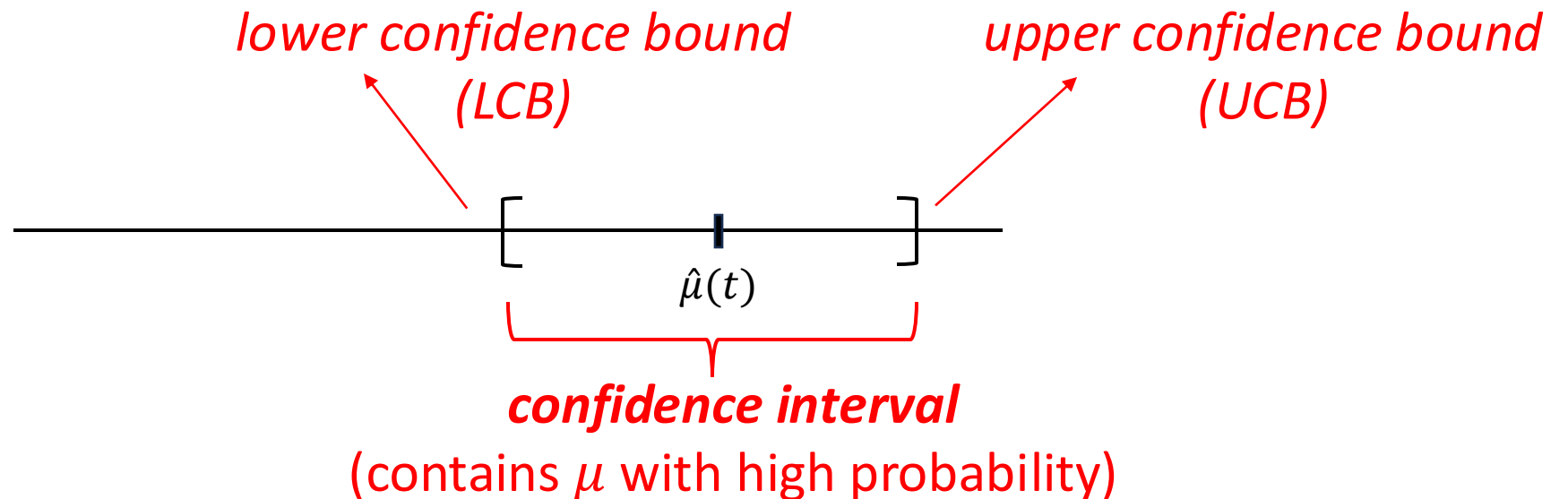
$$\Rightarrow P\left(|\hat{\mu}(t) - \mu| > \sqrt{\frac{2 \log(2/\delta_t)}{t}}\right) \leq \delta_t$$

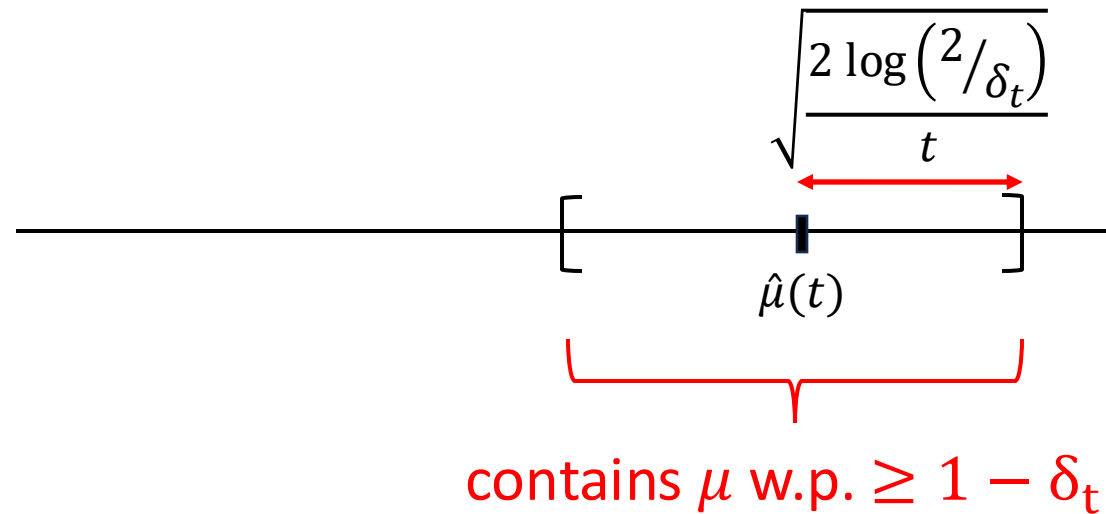


SubGaussian concentration inequality:

$$P(|\hat{\mu}(t) - \mu| > \epsilon) \leq 2 \exp\left(-\frac{t\epsilon^2}{2}\right)$$

$$\Rightarrow P\left(|\hat{\mu}(t) - \mu| > \sqrt{\frac{2 \log(2/\delta_t)}{t}}\right) \leq \delta_t$$





Vanilla approach:

$$\text{Set } \delta_t^i = \frac{\delta}{2Kt^2}$$

$\Rightarrow P(\text{Confidence intervals ever become 'invalid'})$

$$\leq \sum_{i=1}^K \sum_{t=1}^{\infty} \delta_t^i \leq \delta$$

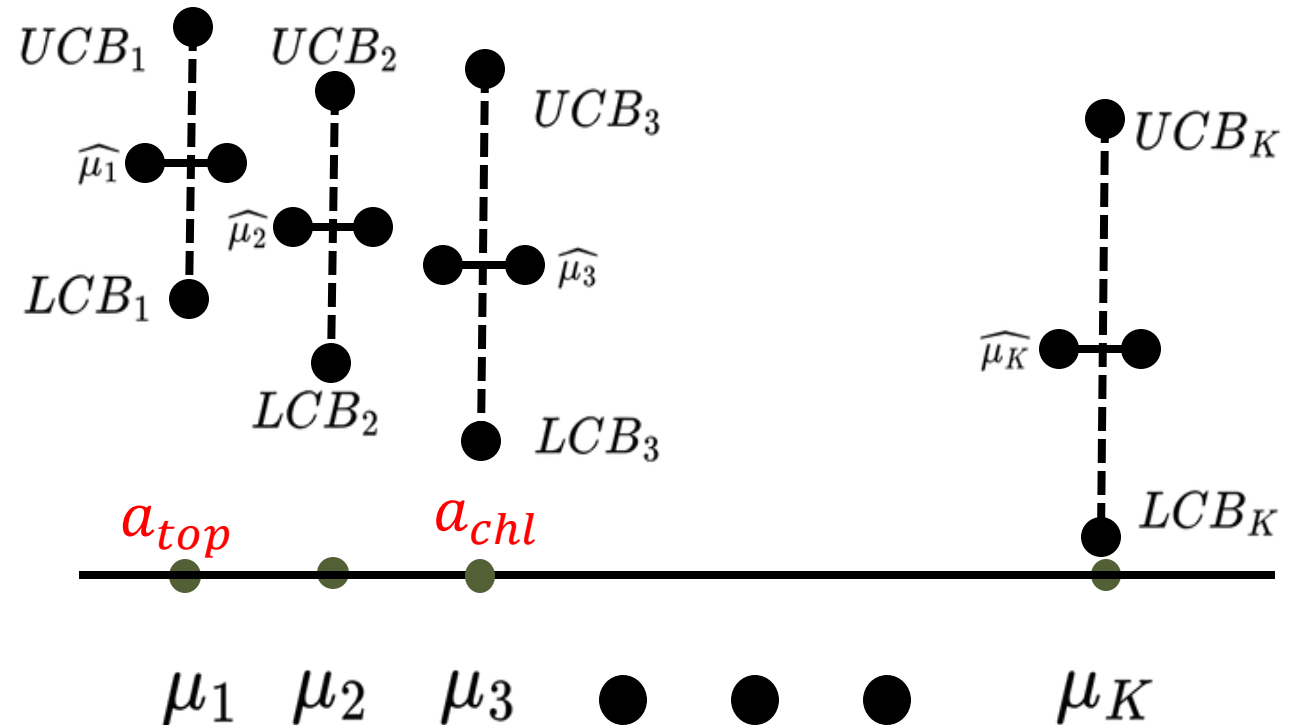
\Rightarrow **All** confidence intervals remain valid **at all times** w.p. $\geq 1 - \delta$

Similar approach works with other arm distribution families

The LUCB algorithm [Kalyanakrishnan et. al, 2013]

- In each round, sample two arms:

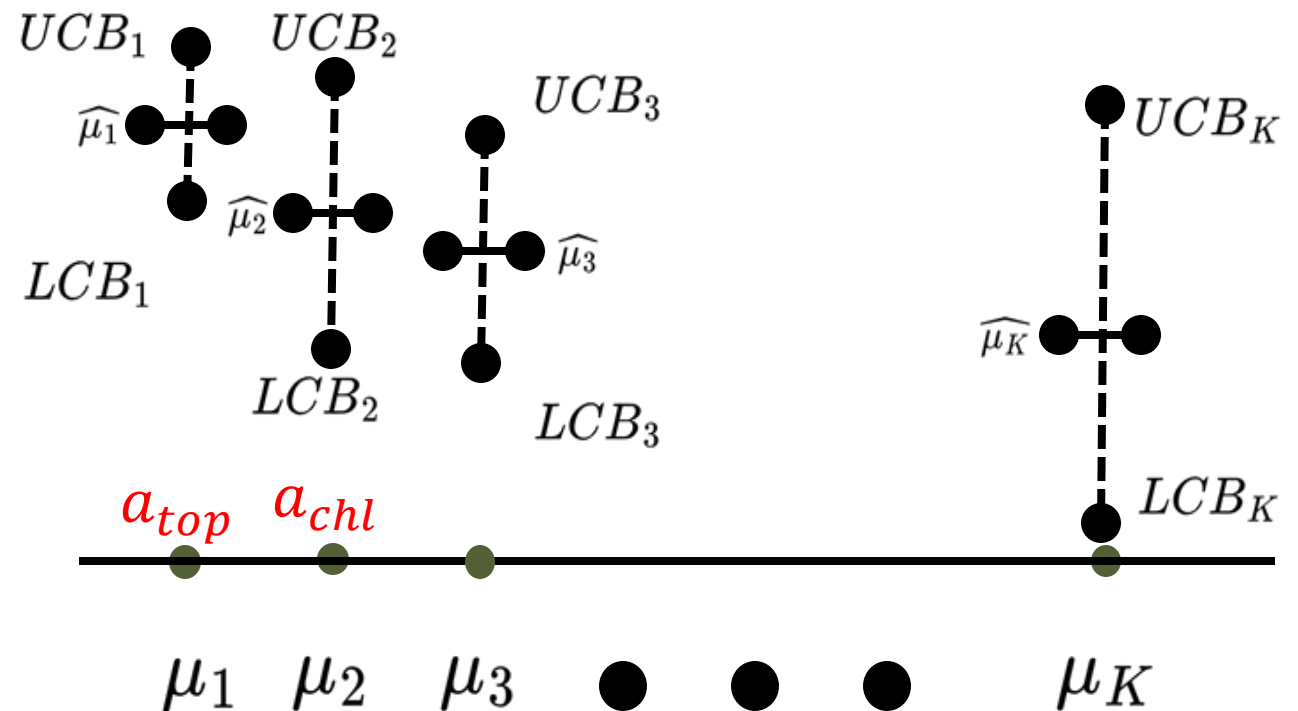
$$a_{top} = \operatorname{argmax}_i \hat{\mu}_i \text{ \& \; } a_{chl} = \operatorname{argmax}_{i \neq a_{top}} UCB_i$$



The LUCB algorithm [Kalyanakrishnan et. al, 2013]

- In each round, sample two arms:

$$a_{top} = \operatorname{argmax} \hat{\mu}_i \text{ \& } a_{chl} = \operatorname{argmax}_{i \neq a_{top}} UCB_i$$

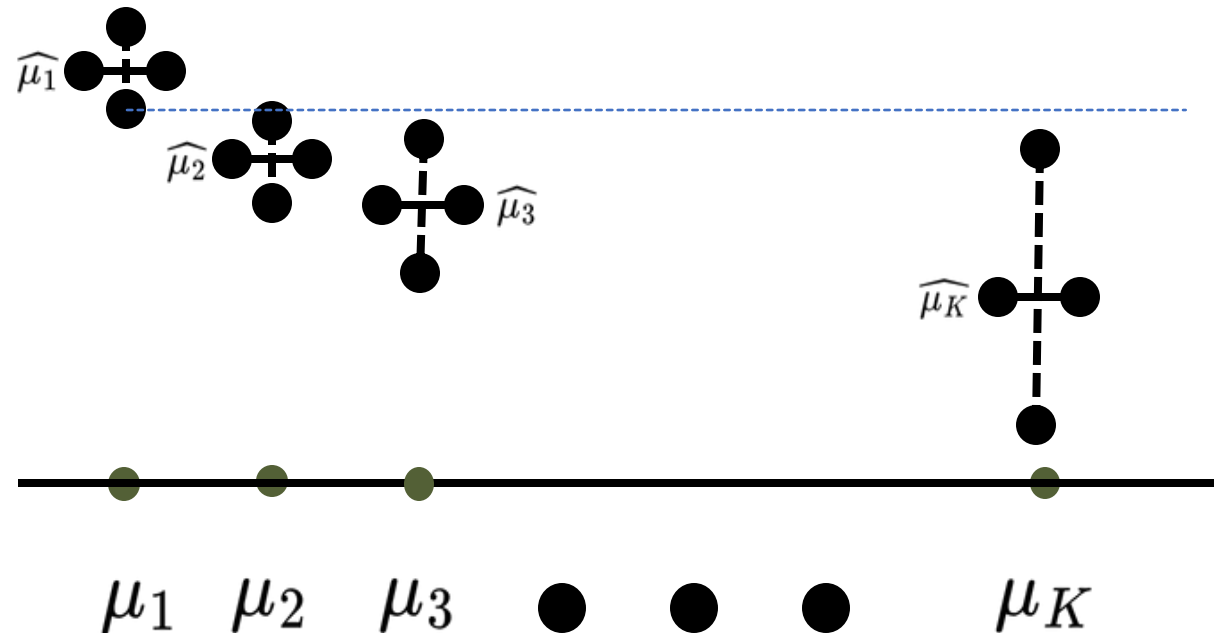


The LUCB algorithm [Kalyanakrishnan et. al, 2013]

- In each round, sample two arms:

$$a_{top} = \operatorname{argmax}_i \hat{\mu}_i \text{ \& \; } a_{chl} = \operatorname{argmax}_{i \neq a_{top}} UCB_i$$

- Stop sampling when $LCB_{a_{top}} > UCB_j$ for all $j \neq a_{top}$
- Recommend arm a_{top}



The LUCB algorithm [Kalyanakrishnan et. al, 2013]

- LUCB is δ -PC
- With probability $\geq 1 - \delta$, number of pulls prior to stopping is

$$\mathcal{O} \left(\sum_i \frac{1}{\Delta_i^2} \log \left(\frac{K \log(\Delta_i^{-2})}{\delta} \right) \right)$$

where $\Delta_i = \mu_1 - \mu_i$ for $i \neq 1$, $\Delta_1 = \Delta_2$

- Similar bound for the *average* stopping time

Q: How good is this?

Information theoretic lower bound [Kaufmann et. al, 2016]


MAB instance ν

$ALT(\nu)$ = set of instances with best arm different from ν

Then for any δ -PC algorithm,

$$E[\tau_\delta] \geq C(\nu) \log \left(\frac{1}{4\delta} \right),$$

$$C(\nu)^{-1} = \sup_{w \in \Sigma_K} \inf_{\lambda \in ALT(\nu)} \sum_i w_i D(\nu_i, \lambda_i)$$

 *probability simplex*

LUCB vs lower bound

Consider 1-Gaussian instance ν ,

$$2 \left(\sum_i \frac{1}{\Delta_i^2} \right) \leq C(\nu) \leq 4 \left(\sum_i \frac{1}{\Delta_i^2} \right) \\ \Rightarrow E[\tau_\delta] \geq 2 \left(\sum_i \frac{1}{\Delta_i^2} \right) \log \left(\frac{1}{4\delta} \right)$$

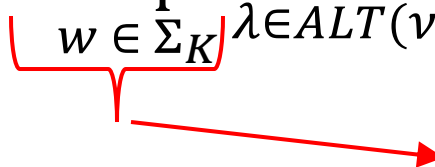
Compare with LUCB bound of $\mathcal{O} \left(\sum_i \frac{1}{\Delta_i^2} \log \left(\frac{K \log(\Delta_i^{-2})}{\delta} \right) \right)$

Matches in loose 'order sense', modulo logarithmic factors

Track & Stop [Kaufmann et. al, 2016]

- Algorithm design motivated by lower bound
- Recall:

$$E[\tau_\delta] \geq C(\nu) \log\left(\frac{1}{4\delta}\right),$$

$$C(\nu)^{-1} = \sup_{w \in \Sigma_K} \inf_{\lambda \in ALT(\nu)} \sum_i w_i D(\nu_i, \lambda_i)$$


Turns out: Optimal pull fractions given by $w^*(\nu)$

T&S: Sample so as to *track* $w^*(\hat{\nu})$ instead;

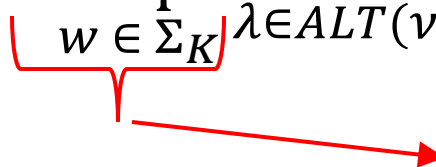
forced exploration (give all arms \sqrt{t} pulls until time t) $\Rightarrow \hat{\nu} \rightarrow \nu$

(works best for parametric distribution families)

Track & Stop [Kaufmann et. al, 2016]

- Algorithm design motivated by lower bound
- Recall:

$$E[\tau_\delta] \geq C(\nu) \log\left(\frac{1}{4\delta}\right),$$

$$C(\nu)^{-1} = \sup_{\substack{w \in \Sigma_K}} \inf_{\lambda \in ALT(\nu)} \sum_i w_i D(\nu_i, \lambda_i)$$


Turns out: Optimal pull fractions given by $w^*(\nu)$

T&S: Sample so as to *track* $w^*(\hat{\nu})$ instead;

forced exploration (give all arms \sqrt{t} pulls until time t) $\Rightarrow \hat{\nu} \rightarrow \nu$

(works best for parametric distribution families)

Track & Stop [Kaufmann et. al, 2016]

$$Z_{i,j}(t) = \log \left[\frac{\max_{\lambda: i > j} \mathcal{L}_{\lambda}(x^t)}{\max_{\lambda: j > i} \mathcal{L}_{\lambda}(x^t)} \right]$$

maximum likelihood
under hypothesis that
arm i beats arm j

maximum likelihood
under hypothesis that
arm j beats arm i

**GLR statistic; captures the extent to which observations
“support” arm i beating arm j**

Stop when $Z_{i,j}(t) > \beta(t, \delta)$ for all $j \neq i$; recommend arm i

(again, works best for parametric distribution families)

Track & Stop [Kaufmann et. al, 2016]



- δ -PC for a suitable choice of $\beta(t, \delta)$
- T&S is known to be asymptotically optimal:

$$\lim_{\delta \rightarrow 0} \frac{E[\tau_{\delta}^{T\&S}]}{\text{Info. theoretic lower bound}} = 1$$

Sampling rule ensures asymptotic optimality (does not depend on δ)

Stopping rule (GLR based) ensures δ -PC

Confidence Intervals v/s T&S

CI based	T&S style
Broadly applicable 	Applicable to parametrized distribution families*
Loose (order sense) stopping time bounds; hard to relate to lower bounds	Explicit interpretable stopping time bounds in asymptotic regime ($\delta \downarrow 0$); asymptotic optimality 

Q: Do confidence interval based algorithms admit explicit & interpretable guarantees in the $\delta \downarrow 0$ regime?

*On the asymptotic optimality of confidence interval based algorithms
for fixed confidence MABs*

Kushal Kejriwal, Nikhil Karamchandani and **J.N.**; AAAI, 2025

LUCB

- In round t , $CI_i = [LCB_i, UCB_i] = [\hat{\mu}_i - r_i, \hat{\mu}_i + r_i]$, where

$$r_i \sim \sqrt{\frac{\log(1/\delta)}{N_i(t-1)}}$$

- Sampling rule:

Pull arm a if $N_a(t-1) < \sqrt{t}$ (forced exploration)

Else, pull $a_{top} = \operatorname{argmax} \hat{\mu}_i$ & $a_{chl} = \operatorname{argmax}_{i \neq a_{top}} UCB_i$

- Stopping rule:

$LCB_{a_{top}} > UCB_j$ for all $j \neq a_{top}$

- Recommend: a_{top}

LUCB

- In round t , $CI_i = [LCB_i, UCB_i] = [\hat{\mu}_i - r_i, \hat{\mu}_i + r_i]$, where

$$r_i \sim \sqrt{\frac{\log(1/\delta)}{N_i(t-1)}}$$

- Sampling rule:

Pull arm a if $N_a(t-1) < \sqrt{t}$ (forced exploration)

*included for
analytical simplicity*

Else, pull $a_{top} = \operatorname{argmax} \hat{\mu}_i$ & $a_{chl} = \operatorname{argmax}_{i \neq a_{top}} UCB_i$

- Stopping rule:

$$LCB_{a_{top}} > UCB_j \text{ for all } j \neq a_{top}$$

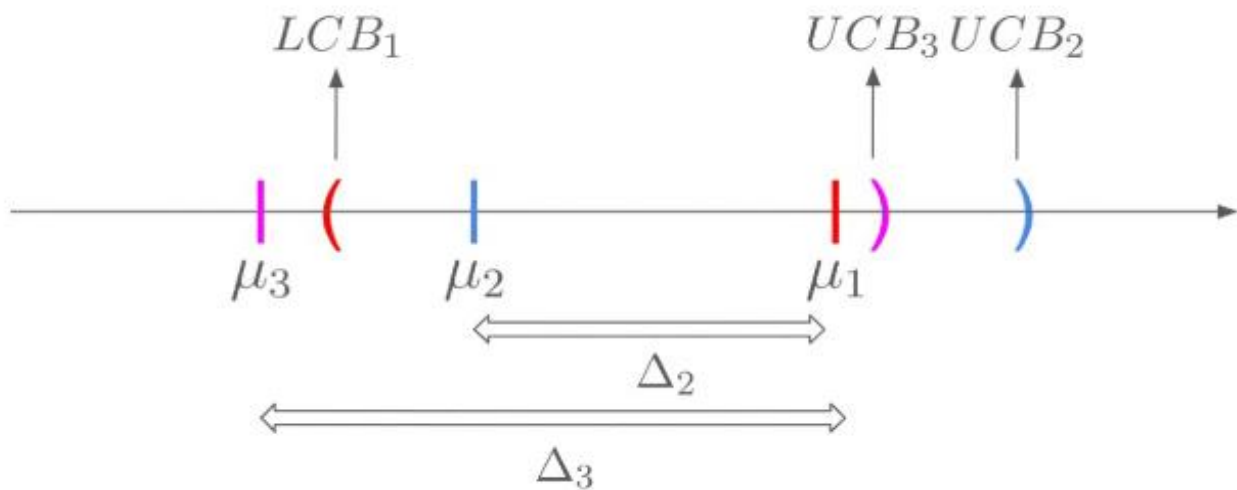
- Recommend: a_{top}

Intuition for the $\delta \downarrow 0$ regime

- Say arm 1 is optimal
- Sample path: Sequence of samples for each arm; can look at different 'copies' of the algorithm running in tandem on same sample path for each value of δ

Note: Sampling process itself is not coupled across δ - contrast with T&S

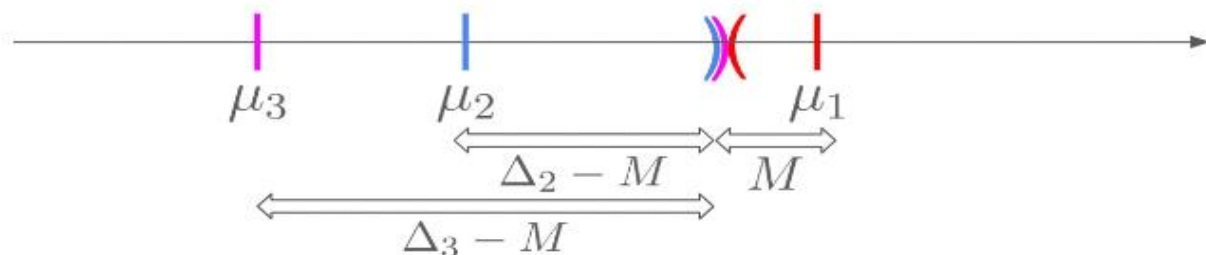
- As $\delta \downarrow 0$, $\tau_\delta \uparrow \infty$, $\hat{\mu}_i \rightarrow \mu$ almost surely (law of large numbers)
- All arm pulls $\propto \log\left(\frac{1}{\delta}\right)$
- Almost surely, after a certain point of time, $a_{top} = 1$



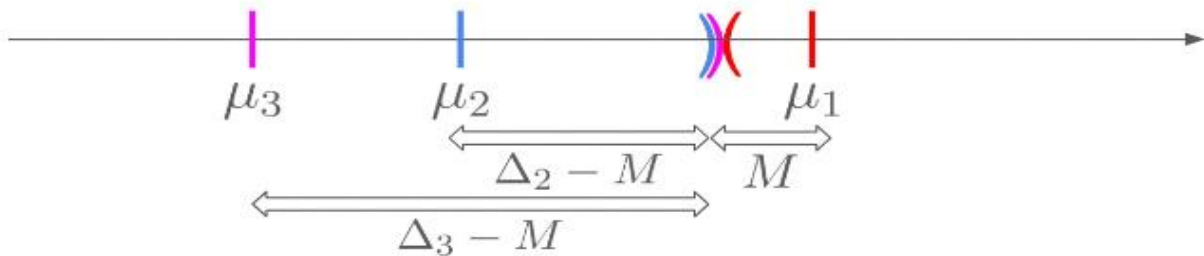
Consider 3 armed instance, $\hat{\mu}_i \approx \mu_i \quad \forall i$
 As we sample,
 $LCB_1 \uparrow$
 $UCB_{a_{chl}} \uparrow$
 a_{chl} alternates between non-optimal arms



UCBs of non-optimal arms align,
 decrease in sync
 LCB of optimal arm \uparrow



At termination,
 $LCB_1 \approx UCB_i \quad \forall i \neq 1$

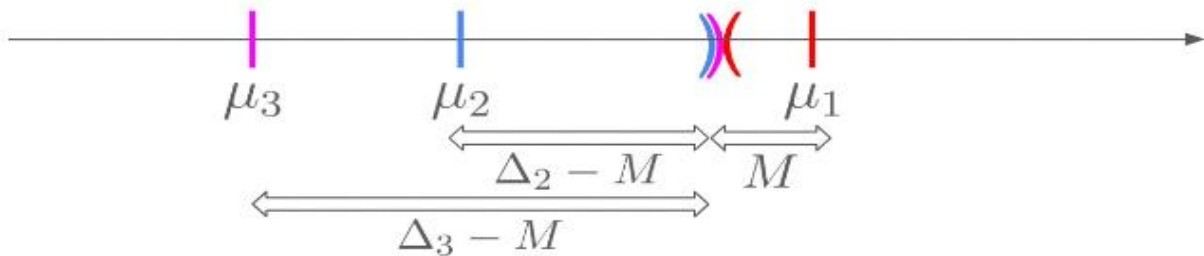


Separation distance M (asymptotically)
invariant as $\delta \downarrow 0$

$$N_1(\tau_\delta) \approx \frac{2}{M^2} \log\left(\frac{1}{\delta}\right) \approx \frac{\tau_\delta}{2}$$

$$N_i(\tau_\delta) \approx \frac{2}{(\Delta_i - M)^2} \log\left(\frac{1}{\delta}\right) \quad \text{for } i \neq 1$$

$$(\text{radius of CI} \sim \sqrt{\frac{\log(1/\delta)}{N_i(t-1)}})$$



Separation distance M (asymptotically)
invariant as $\delta \downarrow 0$

$$N_1(\tau_\delta) \approx \frac{2}{M^2} \log\left(\frac{1}{\delta}\right) \approx \frac{\tau_\delta}{2}$$

$$\tau(\delta) \sim \frac{4}{M^2} \log\left(\frac{1}{\delta}\right)$$

$$N_i(\tau_\delta) \approx \frac{2}{(\Delta_i - M)^2} \log\left(\frac{1}{\delta}\right) \quad \text{for } i \neq 1$$

$$\frac{1}{M^2} = \sum_{i=2}^K \frac{1}{(\Delta_i - M)^2}$$

Theorem: Under LUCB, almost surely,

$$\limsup_{\delta \rightarrow 0} \frac{t(\delta)}{\log(1/\delta)} \leq \frac{4}{M^2}, \text{ where } \frac{1}{M^2} = \sum_{i=2}^K \left(\frac{1}{\Delta_i - M} \right)^2$$

Additionally,

$$\lim_{\delta \rightarrow 0} \frac{N_j(t(\delta))}{t(\delta)} = \begin{cases} \frac{1}{2} & j = 1 \\ \frac{1}{2} \left(\frac{M}{\Delta_j - M} \right)^2 & j \neq 1 \end{cases}$$

Same scaling for expected stopping time as well

Corollary: Under LUCB,

$$\limsup_{\delta \rightarrow 0} \frac{\mathbb{E}(t(\delta))}{\log(1/\delta)} \leq 12 \left(\sum_{i=1}^K \frac{1}{\Delta_i^2} \right)$$

Specializing to a 1-Gaussian instance, recall

$$E[\text{Stopping Time}] \geq 2 \left(\sum_i \frac{1}{\Delta_i^2} \right) \log \left(\frac{1}{4 \delta} \right)$$

$\Rightarrow E[\tau(\delta)] \leq 6$ (Info. Theoretic lower bound)

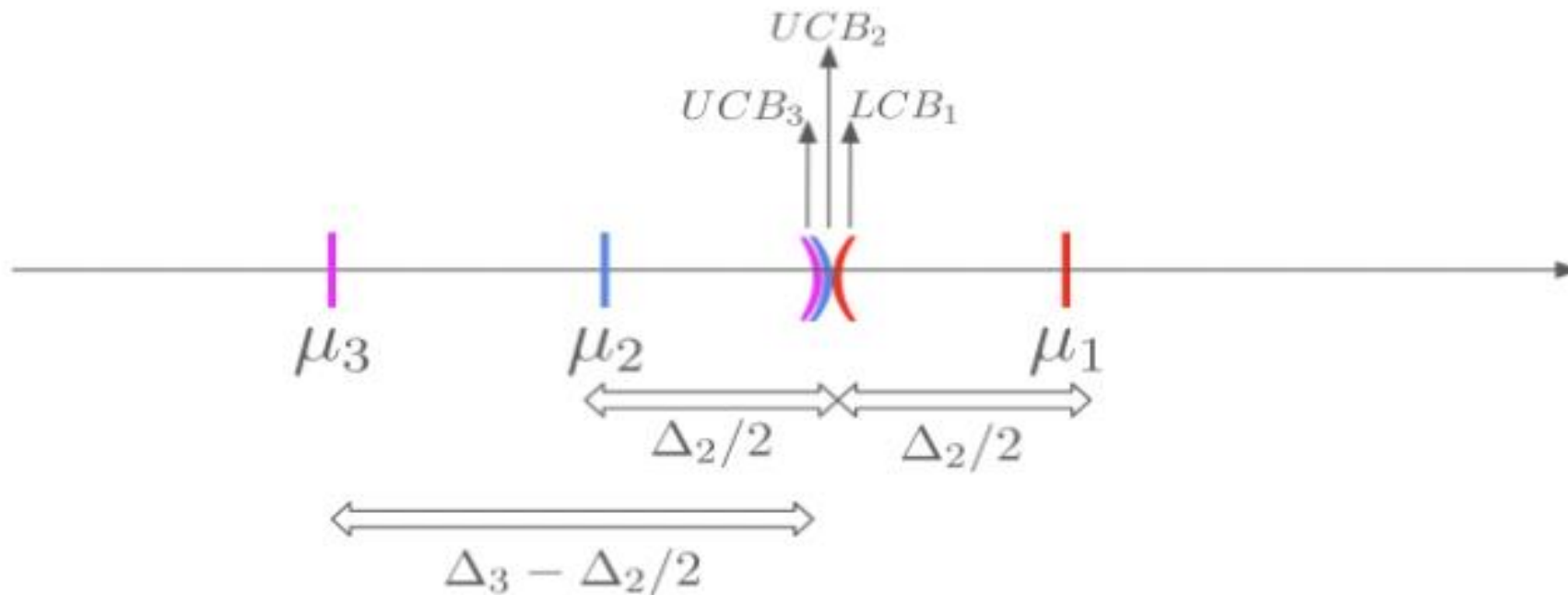
LUCB-Greedy

- Variant of LUCB
- Instead of sampling both a_{top} & a_{chl} ,
sample the one that most shrinks gap between $LCB_{a_{top}}$ and $UCB_{a_{chl}}$
=> sample arm with fewer pulls among a_{top} & a_{chl}

LUCB-Greedy

Analysis similar to that for LUCB

Difference lies in (asymptotic) point of separation between LCBs and UCBs



Theorem: Under LUCB-Greedy, almost surely,

$$\limsup_{\delta \rightarrow 0} \frac{t(\delta)}{\log(1/\delta)} \leq 2M_g, \text{ where } M_g := \left(\frac{8}{\Delta_2^2} + \sum_{i=3}^K \frac{1}{\left(\Delta_i - \frac{\Delta_2}{2}\right)^2} \right)$$

Same scaling for expected stopping time

For 1-Gaussian instances,

$$E[\tau(\delta)] \leq 4 \text{ (Information Theoretic lower bound)}$$

Neither of the upper bounds (for LUCB and LUCB-Greedy) dominates the other

Concluding remarks

- CI algorithms admit a `fluid' analysis in the asymptotic regime as $\delta \downarrow 0$
- Provides a way to better interpret the behavior of these algorithms
- Machinery can be used to analyse/design other CI-based algorithms as well
- Second order analysis for rate of convergence? Finite δ bounds?

Concluding remarks

Asymptotic pull fractions important for (asymptotic) optimality

- Alignment of UCBs of non-optimal arms consistent with lower bound
- Only one relevant degree of freedom : what fraction of pulls to give to optimal arm?
- LUCB *over-samples* optimal arm, LUCB-greedy *under-samples* it
- Can design optimal intermediate optimal algorithm?

Best arm identification in fixed confidence MABs

Confidence intervals and asymptotic optimality

Jayakrishnan Nair

Department of Electrical Engineering, IIT Bombay