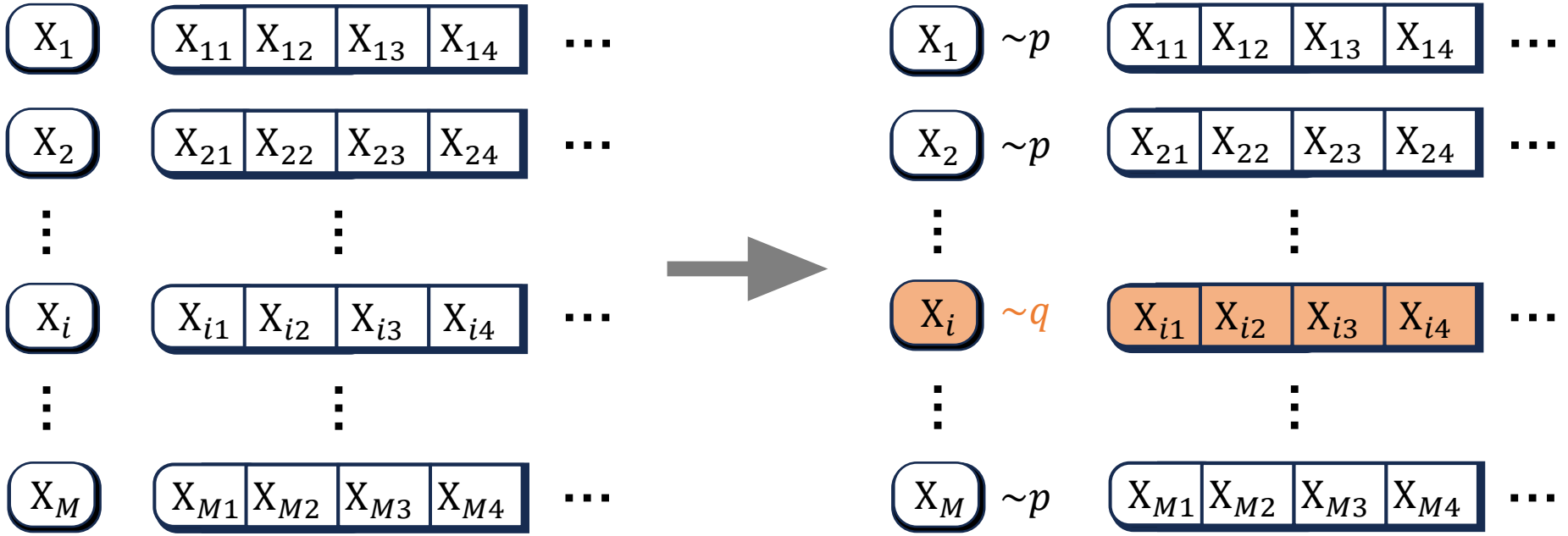
Clustering of Data Sequences

Srikrishna Bhashyam
IIT Madras

Joint work with G. Dhinesh Chandran, Kota Srinivas Reddy

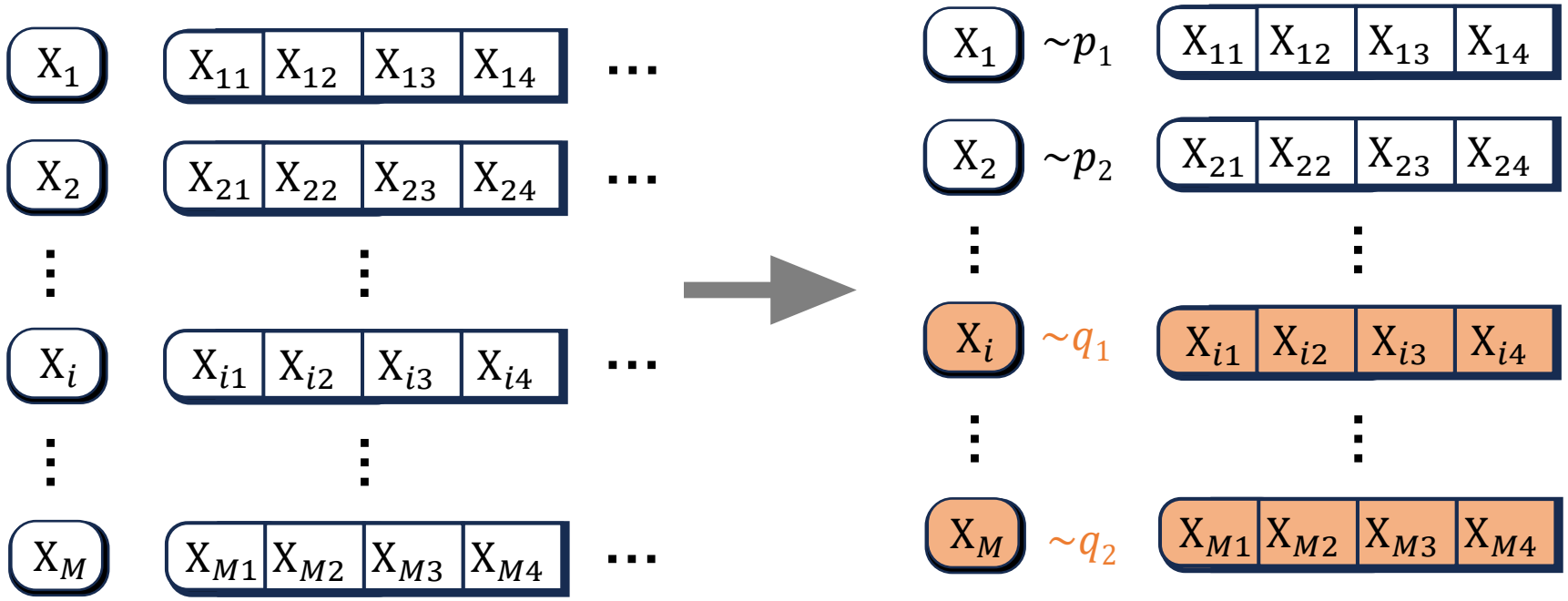
April 13, 2026
IISc Bangalore

Outlying Sequence Detection or Anomaly Detection



- Each data sequence independent and identically distributed (i.i.d.) samples from an unknown distribution
- Typical p vs Anomalous q : Need to find anomalous sequences

Clustering of Data Sequences



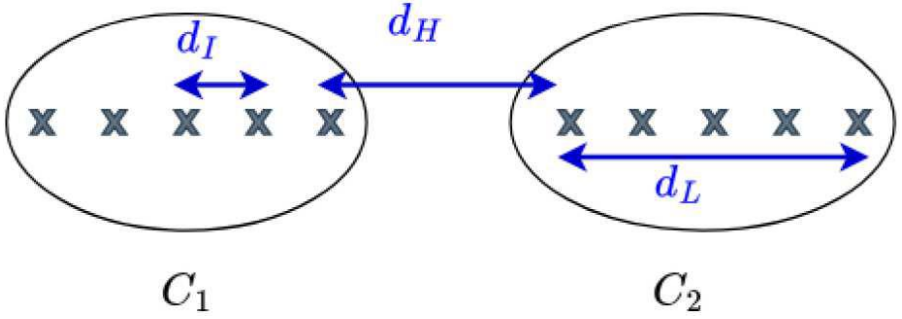
- Each data sequence independent and identically distributed (i.i.d.) samples from an unknown distribution
- Group sequences according to closeness of underlying distributions
- Applications like Network traffic monitoring, market segmentation

Problem Setting: Sample availability

- Fixed Sample Size (FSS): *n* samples from each sequence
  - Decision rule
- Sequential (SEQ): Observe one sample from each sequence at each time
  - Stopping rule and Decision rule
  - Lower expected sample size than FSS for same performance
- Sequential with constraints/Multi-armed bandit setting/Bandit Online Clustering (BOC): Sample one selected sequence at each time
  - Sampling rule, Stopping rule and Decision rule
- Metrics: Clustering error, Sample complexity (Number of samples needed for a given error probability)

Problem Setting: Unknown Distributions

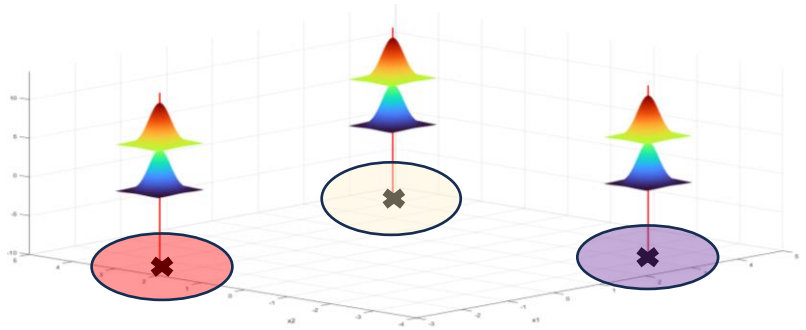
- Class of possible distributions for data sequences
  - Parametric: Only parameter unknown
    - Single-parameter exponential family, Multivariate Gaussian, Sub-Gaussian
  - Nonparametric
- Sequences within a cluster have identical/different distributions



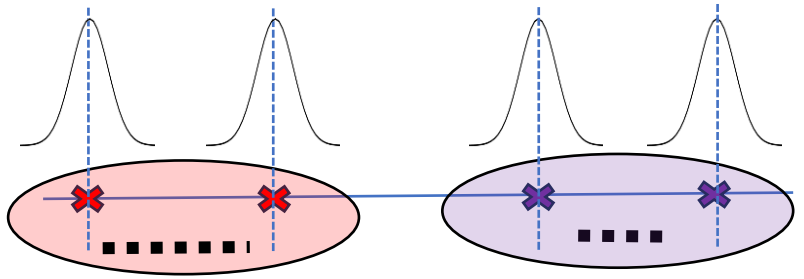
- Clustering method given the distributions/distances
  - k-means, k-medoid, Single Linkage (SLINK), Spectral clustering, etc.

This talk: Bandit Online Clustering (BOC)

	Class of Distributions	Sample dimension	Arms in a cluster	Number of clusters	Clustering method
J. Yang, et al, 2024.	Gaussian	Multi dimensional	Same mean	Multiple clusters	K-means variant
S. Katariya, et al, 2019.	SubGaussian	One dimensional	May have different means	Two clusters	Maximum Gap between means
Our work	**Gaussian/Sub Gaussian/1-parameter exp. Family**	**Multi dimensional**	**May have different means**	**Multiple clusters**	SLINK (Single Linkage Clustering)



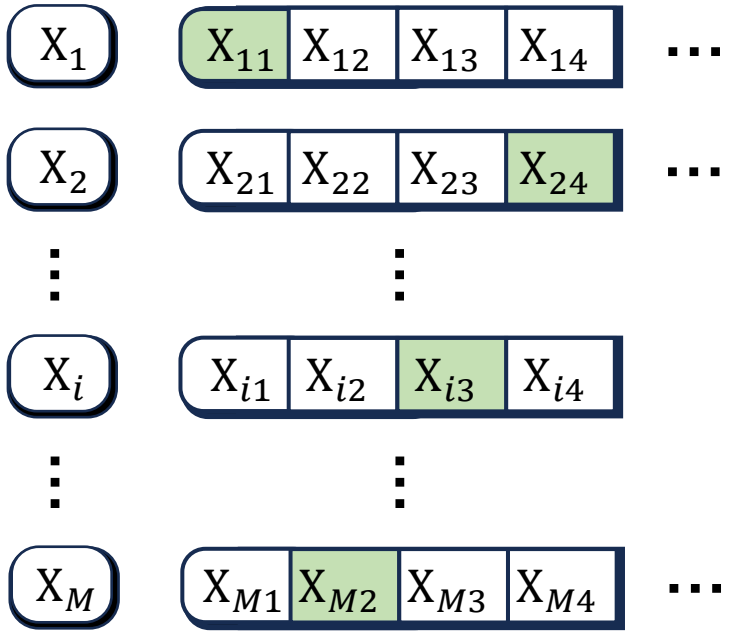
J. Yang, et al, 2024.

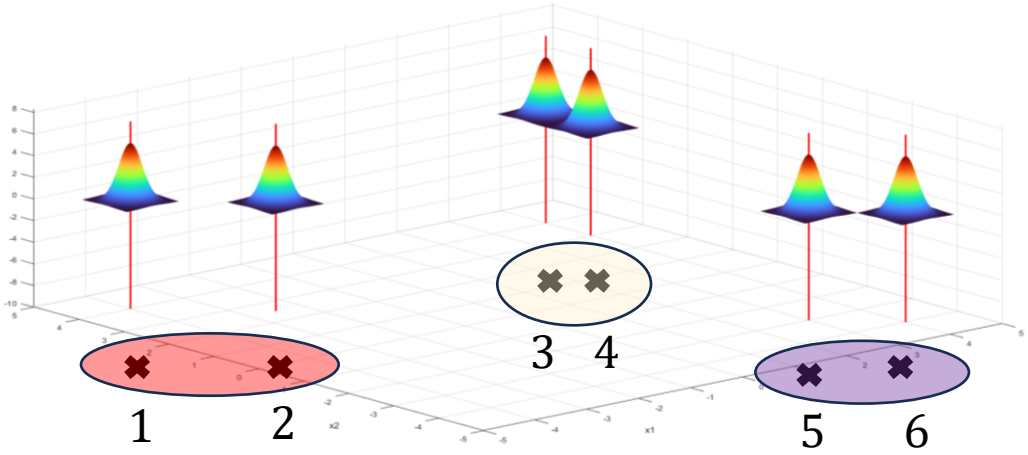


S. Katariya, et al, 2019.

Our Setting

- Multivariate Gaussian, SubGaussian, Single-parameter exponential family
- *M* sequences, *K* clusters





Sampling Rule

Select an arm to get a sample

Stopping Rule

Decides to stop or not

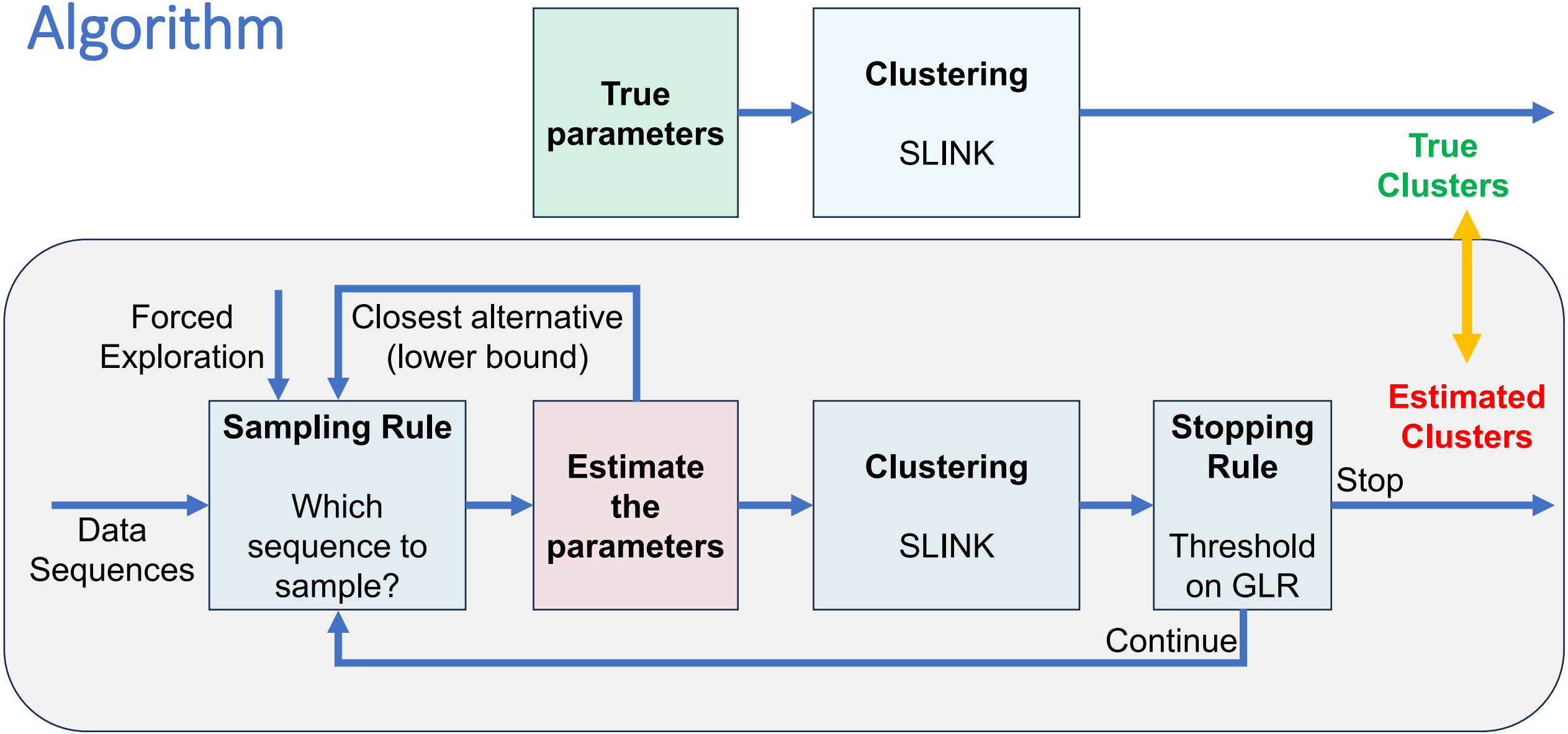
Decision Rule

Estimate clustering of arms

Results

- Average Tracking BOC (ATBOC) algorithm
  - Theoretical upper bound on growth rate of sample complexity vs log(1/Pe)
  - Upper bound is within factor of 2 or 1+γ of the lower bound
- LUCBBOC algorithm
  - Simpler sampling rule based on confidence bounds
  - Good performance with reduced complexity
- Performance improvement over FSS and Sequential
  - Synthetic datasets
  - Real datasets: Traffic monitoring, market segmentation

Algorithm


True parameters
Clustering
SLINK
True Clusters
Forced Exploration
Closest alternative (lower bound)
Sampling Rule
Which sequence to sample?
Estimate the parameters
Clustering
SLINK
Stopping Rule
Threshold on GLR
Estimated Clusters
Stop
Data Sequences
Continue

Closest Alternative: Multivariate Gaussian case

$\boldsymbol{w} = [w_1, w_2, \dots, w_i, \dots, w_M]$ (Sample proportions)

$$w_i = \frac{number\ of\ samples\ from\ i^{th}\ sequence}{total\ number\ of\ samples\ from\ all\ sequences}$$

Closest Alternative

$$\psi(\boldsymbol{w}, \boldsymbol{\mu}) = \inf_{\boldsymbol{\lambda} \in Alt(\boldsymbol{\mu})} \frac{1}{2} \sum_{m=1}^{M} w_m \|\boldsymbol{\lambda}_m - \boldsymbol{\mu}_m\|^2$$

Weighted Distance Between true mean and its closest alternative

Bigger the distance, easier the identification of true clusters

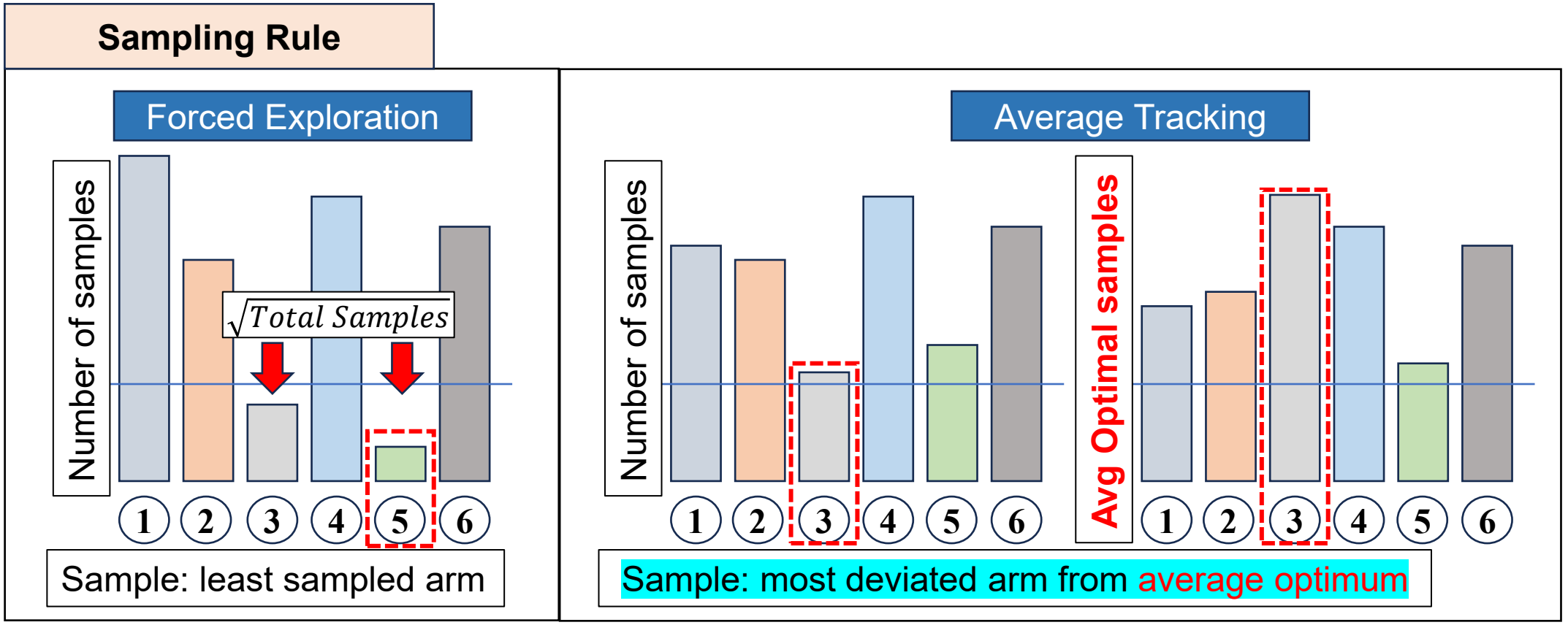
$$\sup_{\boldsymbol{w}} \psi(\boldsymbol{w}, \boldsymbol{\mu})$$

Optimal sampling proportions

$\boldsymbol{\mu}$ - True mean vectors of all arms

$\boldsymbol{\lambda}$ – an alternative to $\boldsymbol{\mu}$
Mean vectors of all arms which forms different cluster than that of $\boldsymbol{\mu}$

Sampling Rule: Average Tracking



$w(s)$ - Arm pull proportions that maximizes the distance between the empirical mean and its closest alternative at time s.

Optimum

$$w(s) \in \arg\max_{\boldsymbol{w} \in P_M} \psi(\boldsymbol{w}, \hat{\boldsymbol{\mu}}(s))$$

Average Optimal proportion

$$\frac{\sum_{s=1}^{t} \boldsymbol{w}(\boldsymbol{s})}{t}$$

Stopping Rule and Decision Rule

Stopping Rule

$$\psi\left(\frac{\boldsymbol{N}(t)}{t}, \widehat{\boldsymbol{\mu}}(t)\right) \geq \text{threshold}$$

Distance between the empirical means and its closest alternative.

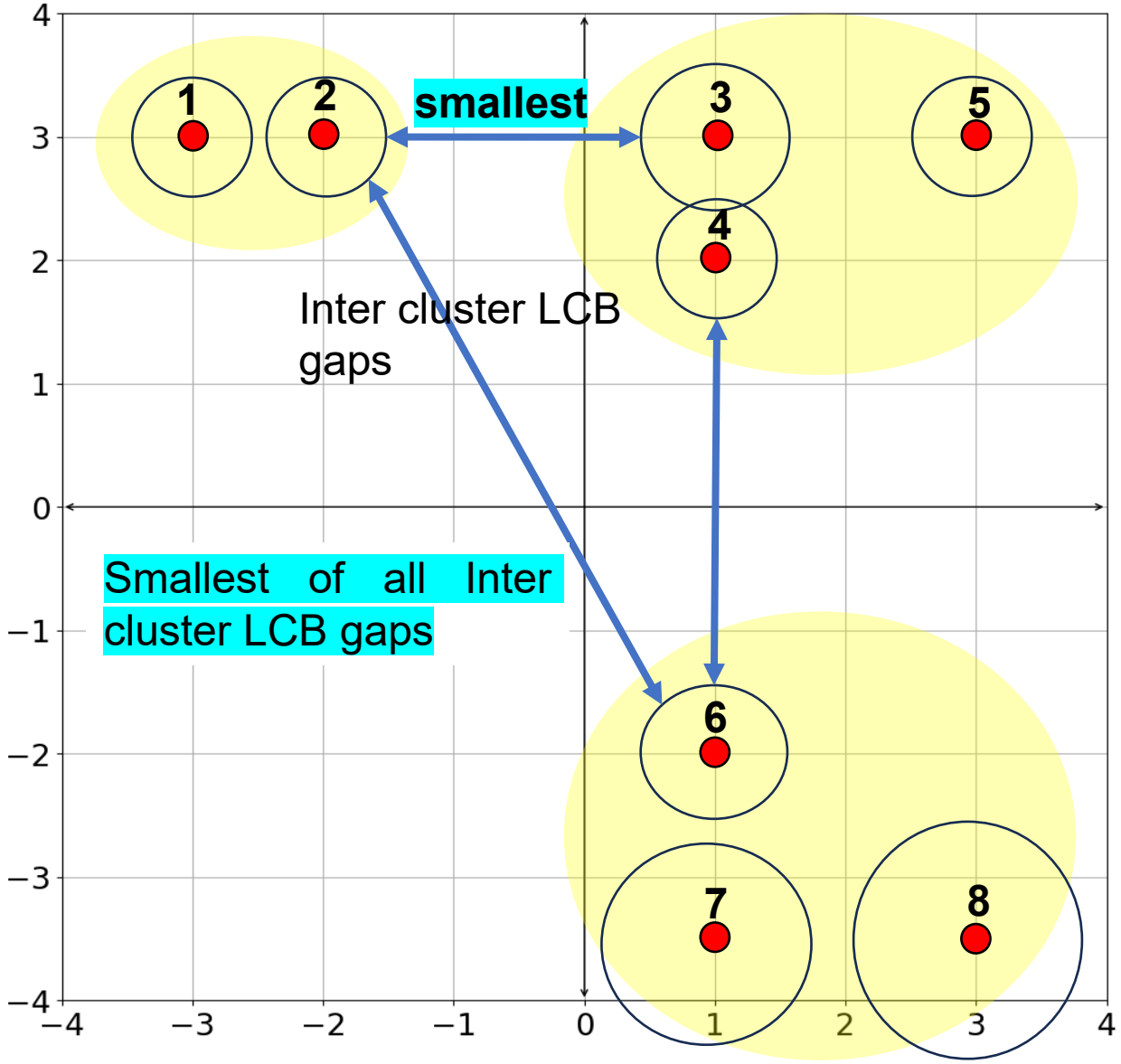
→ Stop

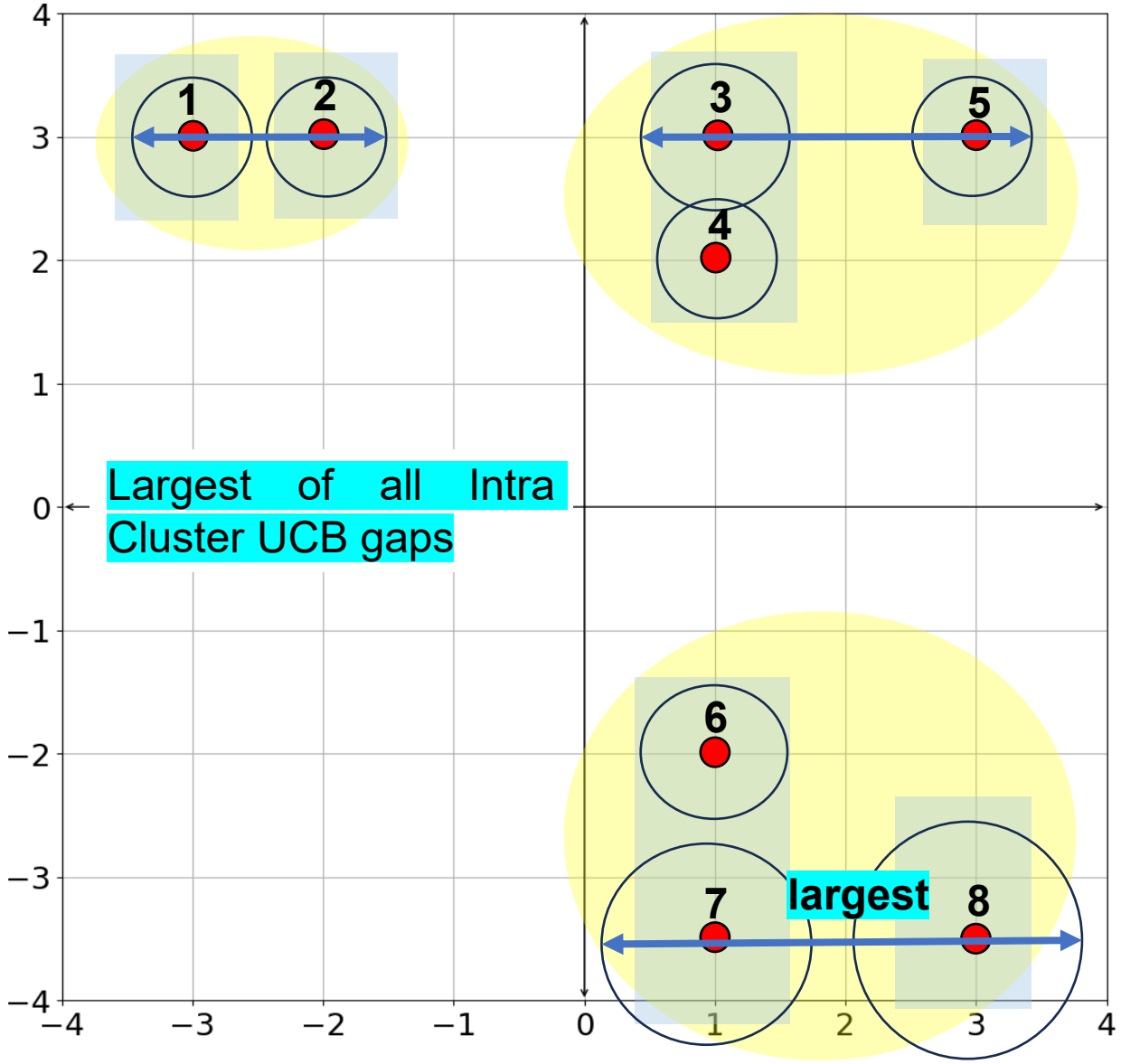
else → Continue

Decision Rule

Single-Linkage (SLINK) clustering on empirical mean vector $\widehat{\boldsymbol{\mu}}$

Sampling Rule: LUCB-based

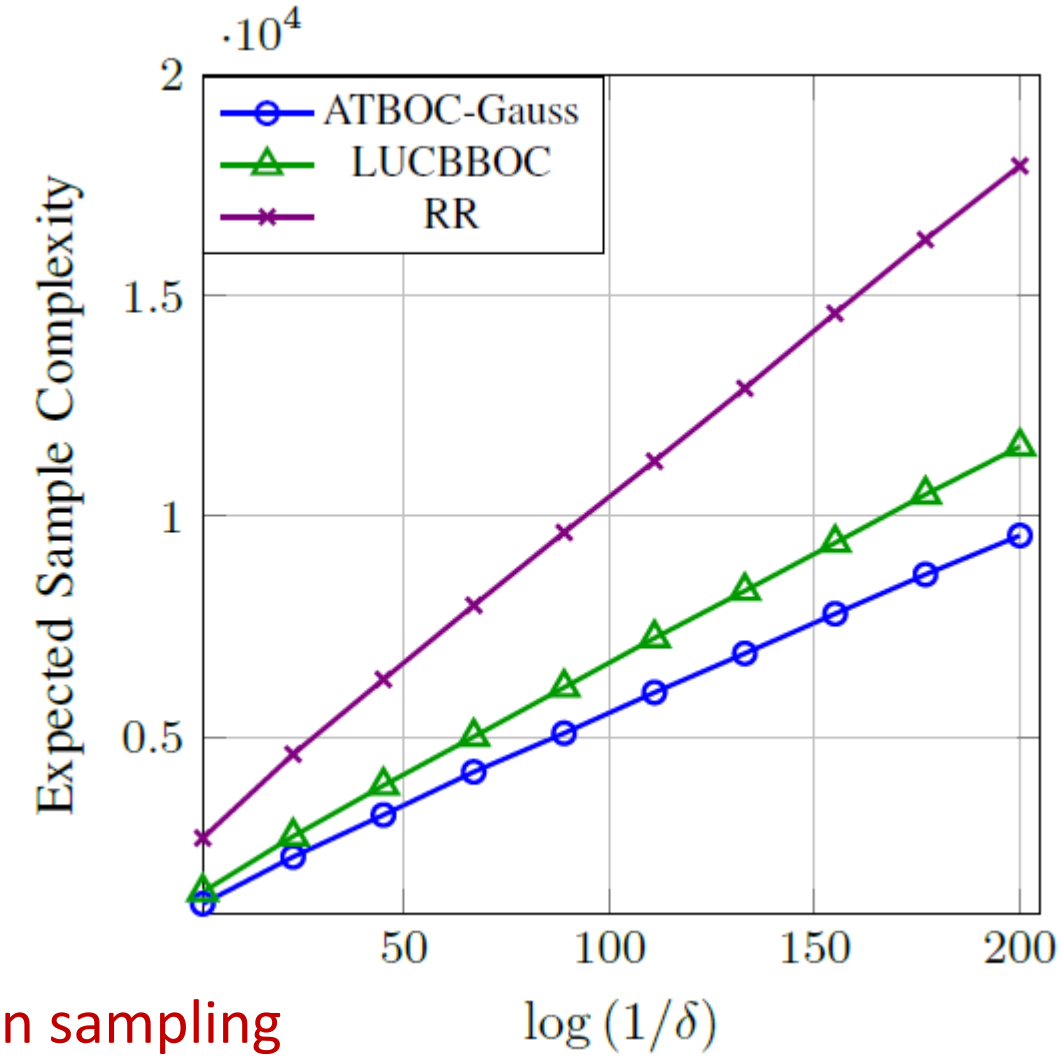




Numerical Results: Synthetic dataset

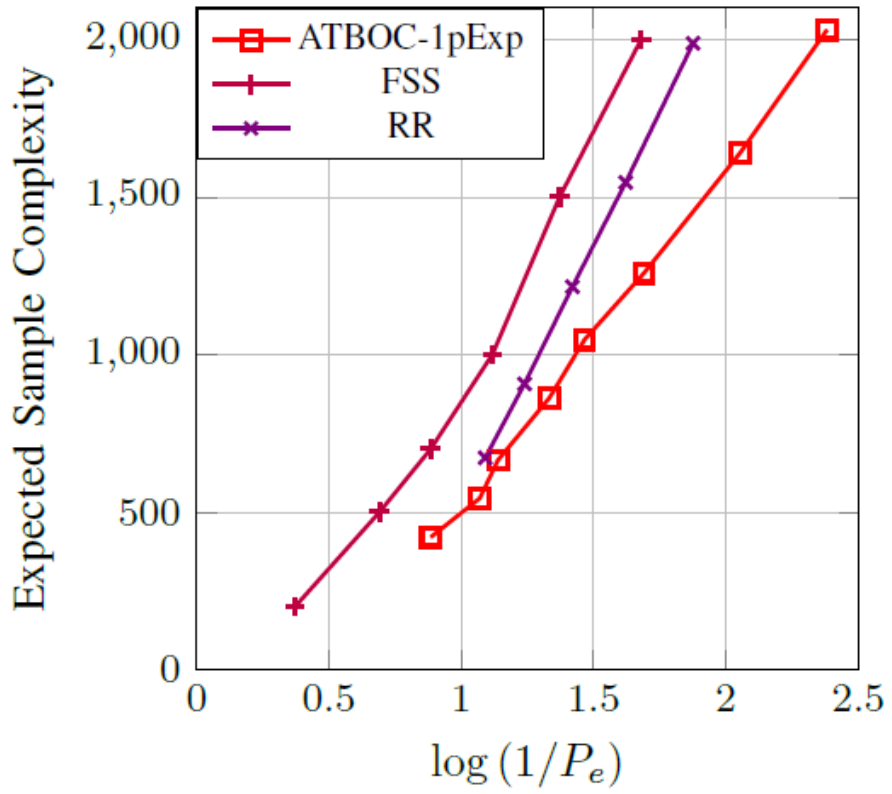
- M = 6 data sequences
- K = 3 clusters
- 2-dimensional Gaussian samples

	Slope	Slope Upper Bound
ATBOC	40	40
LUCBBOC	48	48
RR	75	73

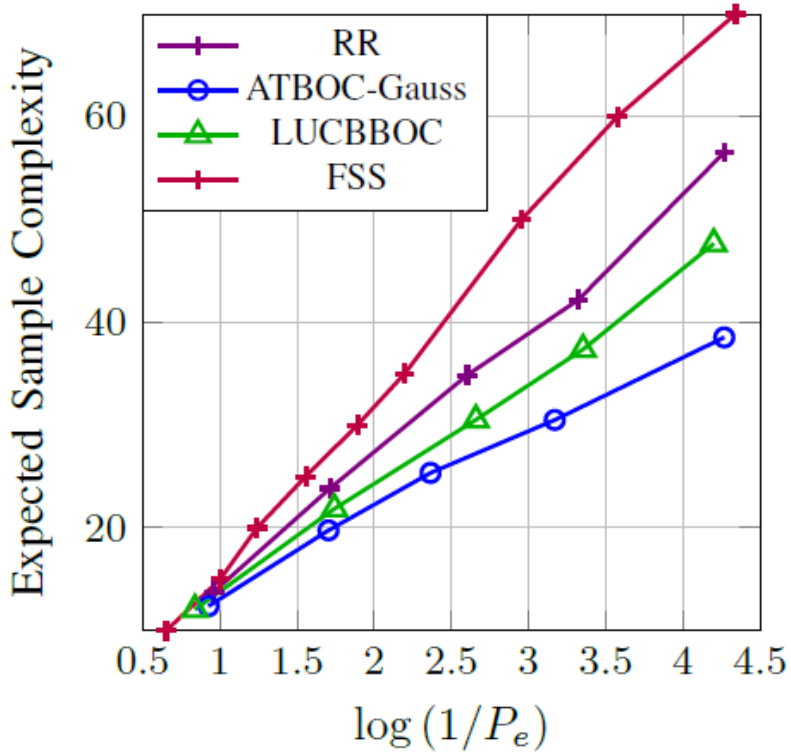


- Significant improvement over round robin sampling

Numerical Results: Real datasets



NYC TLC Dataset: Traffic arrival data and clustering of 6 regions
6 sequences, 3 clusters



MovieLensDataset: 2-D Ratings data and clustering of 6 users
6 sequences, 4 clusters

- Sequential policy significantly better than FSS
- Significant improvement over round robin sampling

Summary

- Average Tracking BOC (ATBOC) algorithm
  - Theoretical upper bound on growth rate of sample complexity vs log(1/Pe)
  - Upper bound is within factor of 2 or 1+γ of the lower bound
- LUCBBOC algorithm
  - Simpler sampling rule based on confidence bounds
  - Good performance with reduced complexity
- Performance improvement over FSS and Sequential
  - Synthetic datasets
  - Real datasets: Traffic monitoring, market segmentation

- For more details and extensions
  - G. D. Chandran, K. S. Reddy, and S. Bhashyam, “Online clustering with bandit information,” in 2025 IEEE International Symposium on Information Theory (ISIT), Ann Arbor, MI, USA, 2025.
  - G. D. Chandran, K. S. Reddy, S. Bhashyam, Asymptotically Optimal Bandit Online Clustering for Single Parameter Exponential Family of Distributions, 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, May 2026.
  - G. D. Chandran, K. S. Reddy, and S. Bhashyam, “Online clustering of data sequences with bandit information,” 2025. [Online]. Available: https://arxiv.org/abs/2501.11421
  - G. D. Chandran, K. S. Reddy, and S. Bhashyam, “Efficient clustering in stochastic bandits,” 2026. [Online]. Available: https://arxiv.org/abs/2601.09162 (Accepted in part ISIT 2026)
  - G. D. Chandran, K. S. Reddy, and S. Bhashyam, “Sequential Spectral Clustering of Data Sequences,” 2025. [Online]. Available: https://arxiv.org/abs/2509.09144 (Accepted in part ISIT 2026)

Thank you

https://www.ee.iitm.ac.in/skrishna/