Exponentially consistent K-means clustering algorithm based on Kolmogrov-Smirnov test

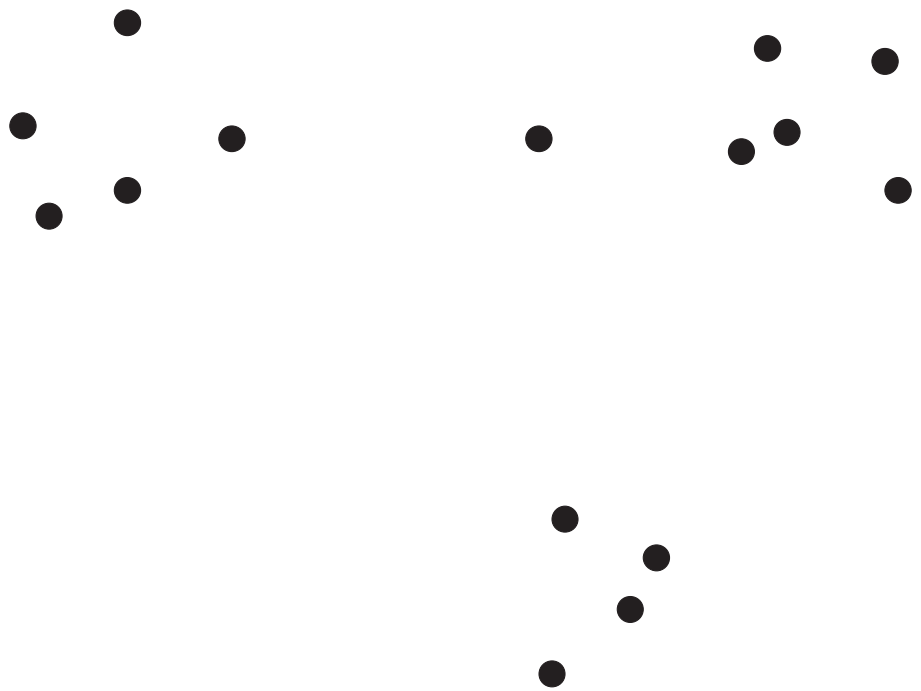
Tiexing Wang*, Donald J. Bucci Jr.†, Yingbin Liang⋆, Biao Chen*, Pramod K Varshney*

*Department of EECS, Syracuse University, Syracuse, NY, 13244, USA
†Lockheed Martin - Advanced Technology Labs, Cherry Hill, NJ, 08002, USA
⋆Department of ECE, The Ohio State University, Columbus, OH 43210, USA

ICASSP 2018

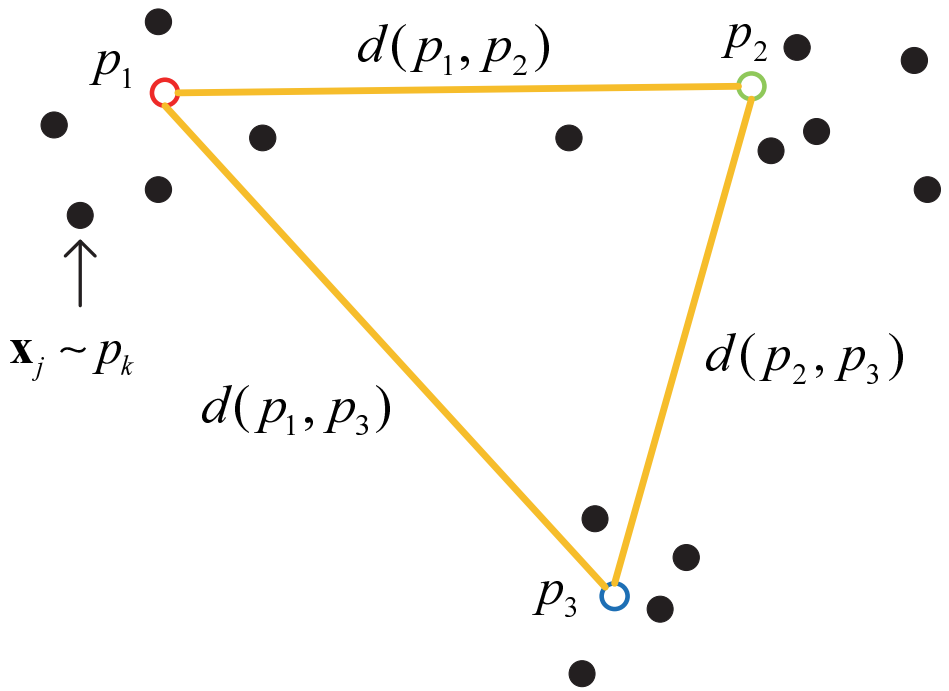
This material is based upon work supported in part by the Defense Advanced Research Projects Agency under Contract No. HR0011-16-C-0135 and by the Dynamic Data Driven Applications Systems (DDDAS) program of AFOSR under grant number FA9550-16-1-0077.

Motivation



- non-parametric
- continuous distribution

Notation



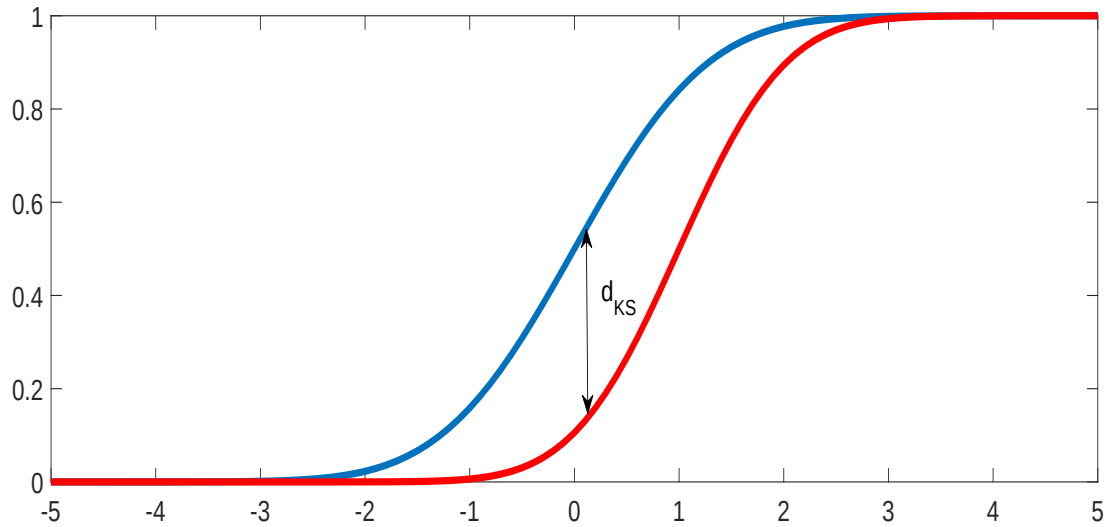
- $K = 3$ - # of distributions
- $M = 15$ - # of sequences

Kolmogrov-Smirnov (KS) distance

- Empirical c.d.f. - $F_{\mathbf{x}}(a) = \frac{1}{n}\sum_{i=1}^{n} 1_{[-\infty,a]} x_i$
  - $\mathbf{x} = \{x_1, \ldots, x_n\}$ - data sequence
  - $1_{[-\infty,x]}$ - indicator function
- KS distance:

$$d_{KS}(\star, *) = \sup_{a \in \mathbb{R}} \left| F_{\star}(a) - F_{*}(a) \right|,$$

  - $\star$, $*$ - sample data or distributions.



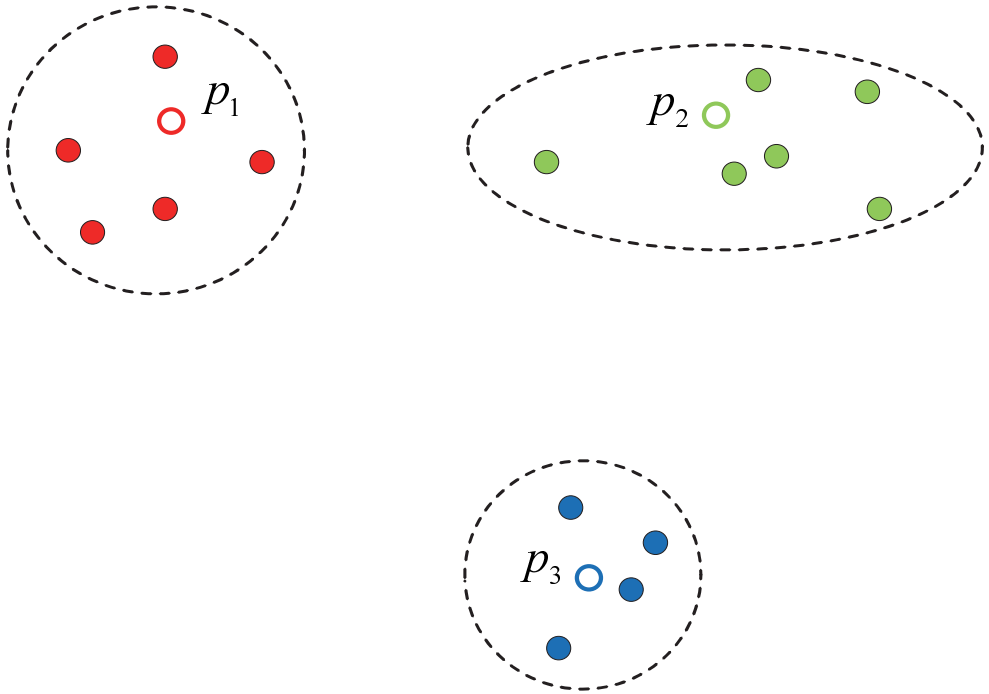
Fundamental lemma

Lemma 1 (Massart1990)

Suppose $\mathbf{x}$ *is generated by* p *and* $F_{\mathbf{x}}(a)$ *is the corresponding empirical c.d.f. Then*

$$P\big(d_{KS}(\mathbf{x}, p) > \epsilon\big) \leq 2\exp\big(-2n\epsilon^2\big).$$

Reasonable clustering result given p_i's



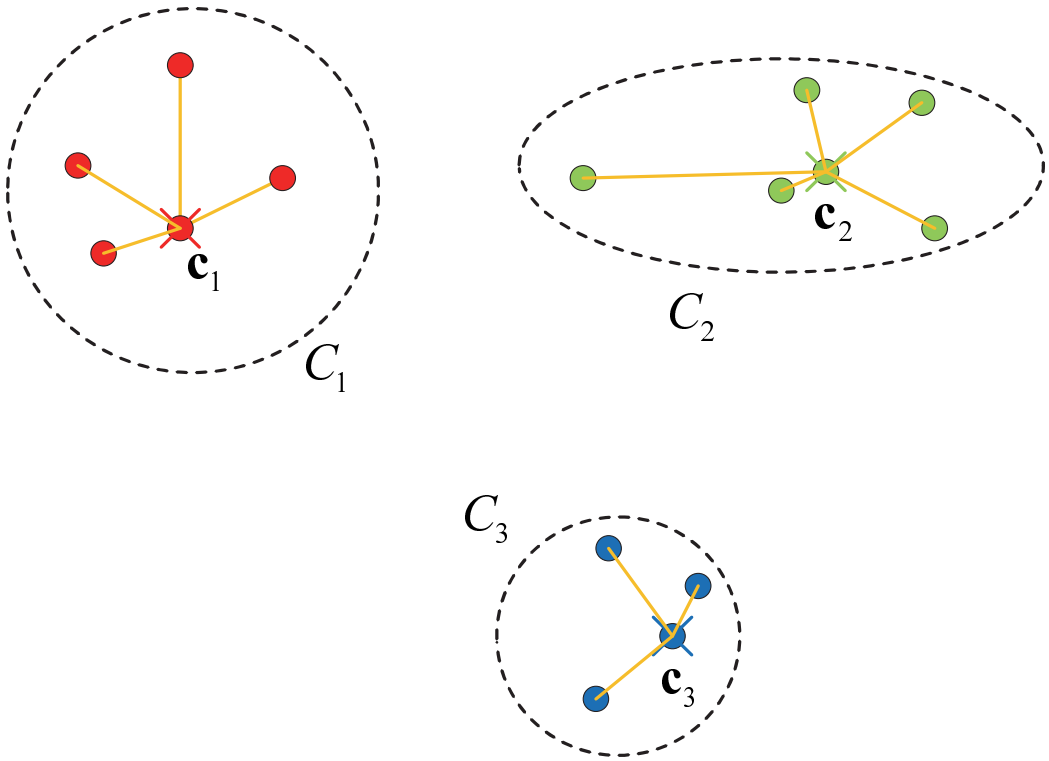
Fundamental lemma cont.

Lemma 2

Suppose $\mathbf{x}$ *and* $\mathbf{z}$ *are generated by* p_1*, and* $\mathbf{y}$ *is generated by* p_2*. Then,*

$$P\left(d_{KS}(\mathbf{x}, \mathbf{z}) > d_{KS}(\mathbf{y}, \mathbf{z}) \right) \leq 6 \exp\left(- \frac{n d_{KS}^2(p_1, p_2)}{8} \right).$$

Reasonable clustering result with unknown p_i's



- Cost function - $J = \sum_{l=1}^{3} \sum_{\mathbf{x}_j \in C_l} d_{KS}(\mathbf{x}_j, \mathbf{c}_l)$

Contributions

- Exponential consistency is established for the proposed algorithm.
  - P_e - the probability of error of a clustering algorithm
  - n - sample size
  - Consistency: $\lim_{n\to\infty} P_e = 0$
  - Exponentially consistency: $\lim_{n\to\infty} -\frac{1}{n} \log P_e > 0$.

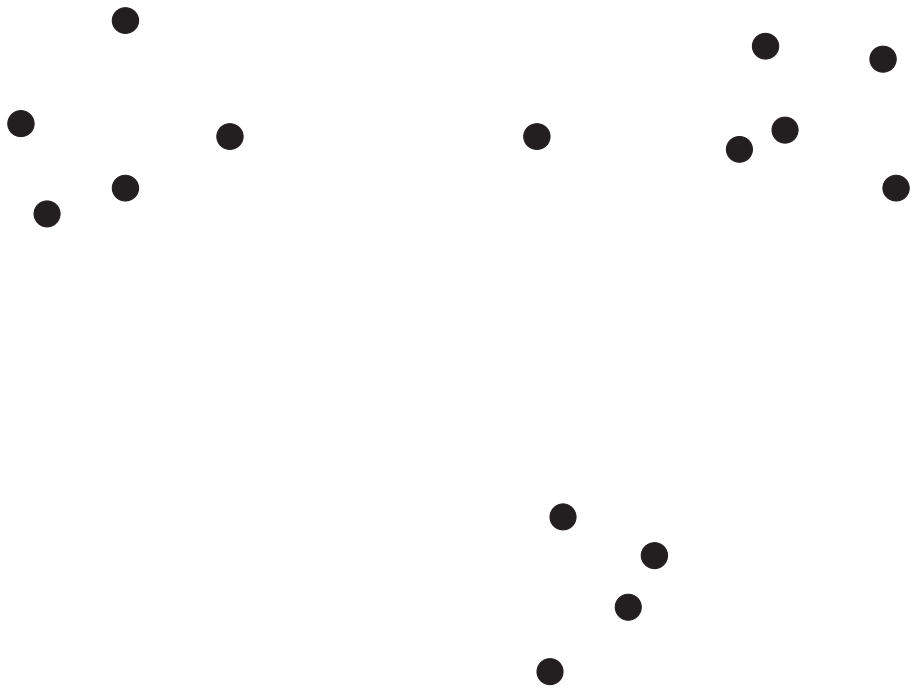
Clustering given K - Initialization

Algorithm 1 KS-based initialization given K for composite distributions

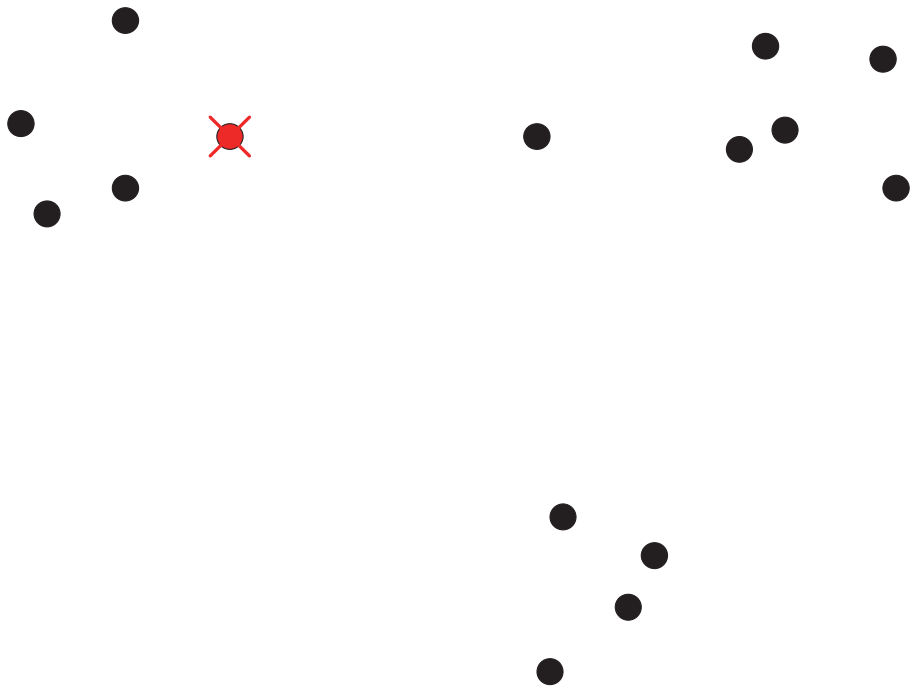
1: **Input**: $\{\mathbf{x}_j\}_{j=1}^{M}$, number of clusters K.
2: {Center initialization}
3: {Cluster initialization}
4: **Output**: $\{\mathcal{C}_k\}_{k=1}^{K}$

- Center initialization
  - Maximize the minimal distance [Katsavounidis-etal 1996]
  - Randomly choose [Moreno-Sáez-etal 2014]

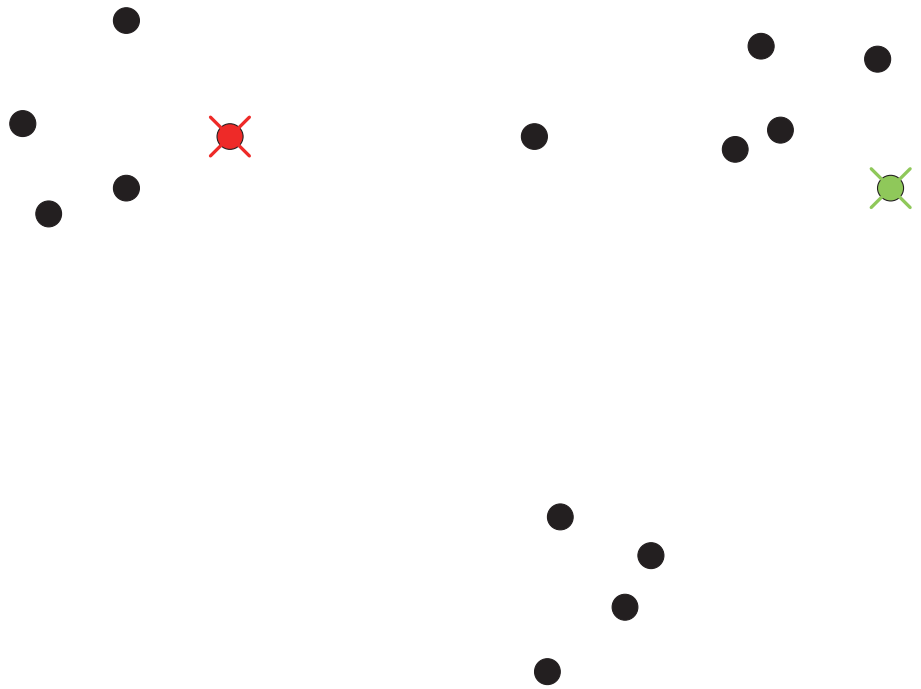
Center initialization illustration



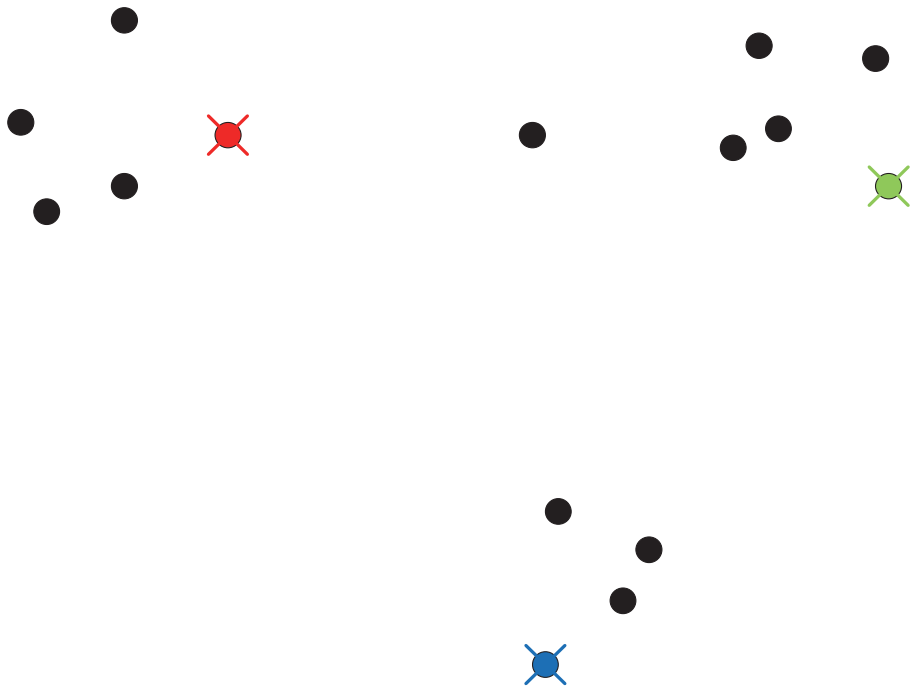
Center initialization illustration cont.



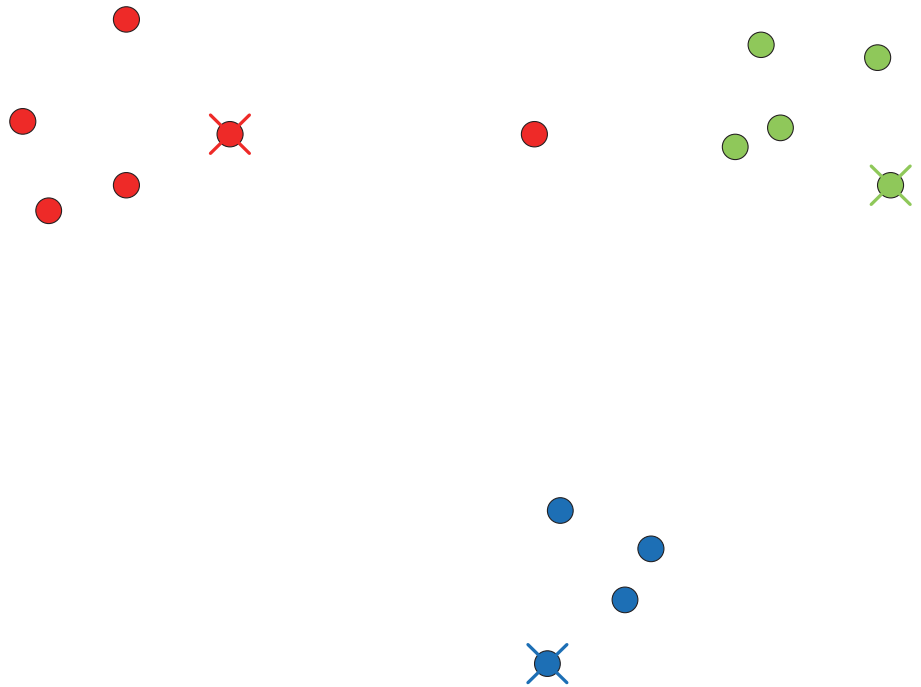
Center initialization illustration cont.



Center initialization illustration cont.



Cluster initialization illustration



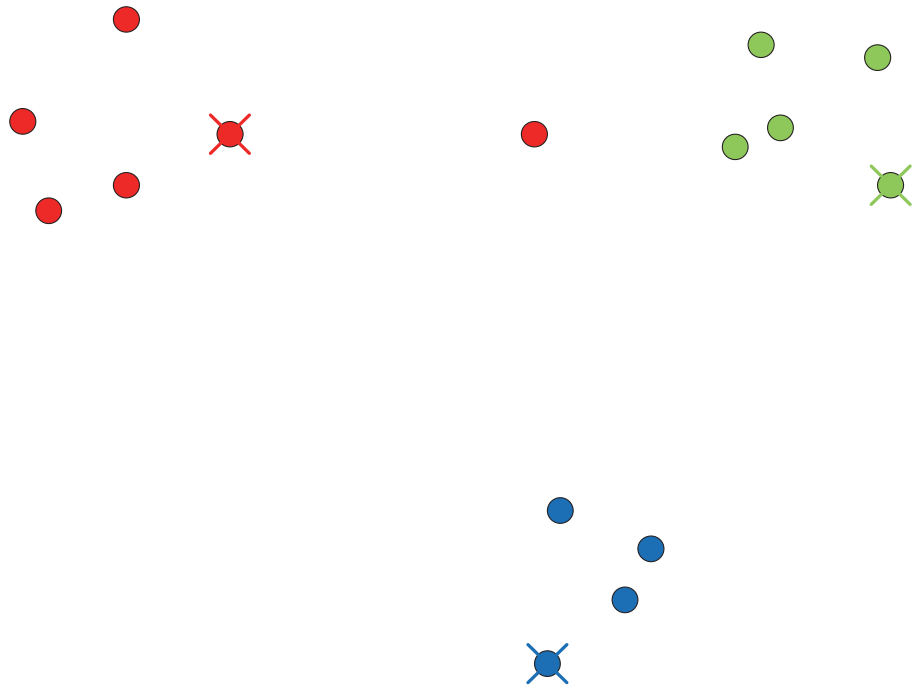
- Cost J_1^0

Clustering given K - Iteration

Algorithm 2 KS based clustering given K for composite distributions

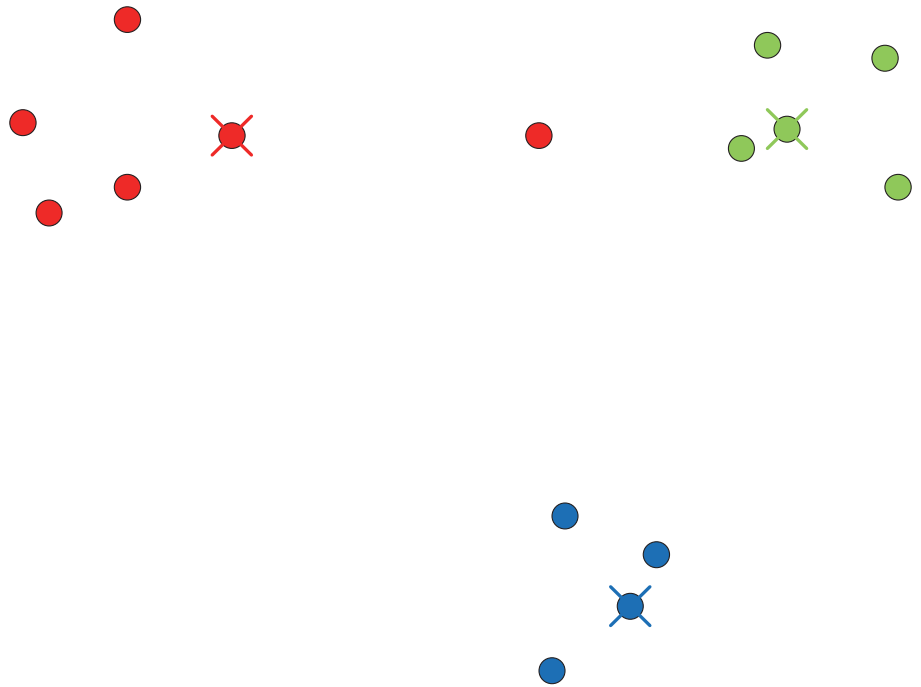
1: **Input**: $\{\mathbf{x}_j\}_{j=1}^{M}$, number of clusters K.
2: **Initialization**: obtain $\{\mathcal{C}_k\}_{k=1}^{K}$ by Algorithm 1.
3: **while** the clustering result does not converge **do**
4: $\quad$ {Center update}
5: $\quad$ {Cluster update}
6: **end while**
7: **Output**: $\{\mathcal{C}_k\}_{k=1}^{K}$.

Iteration one - center update illustration



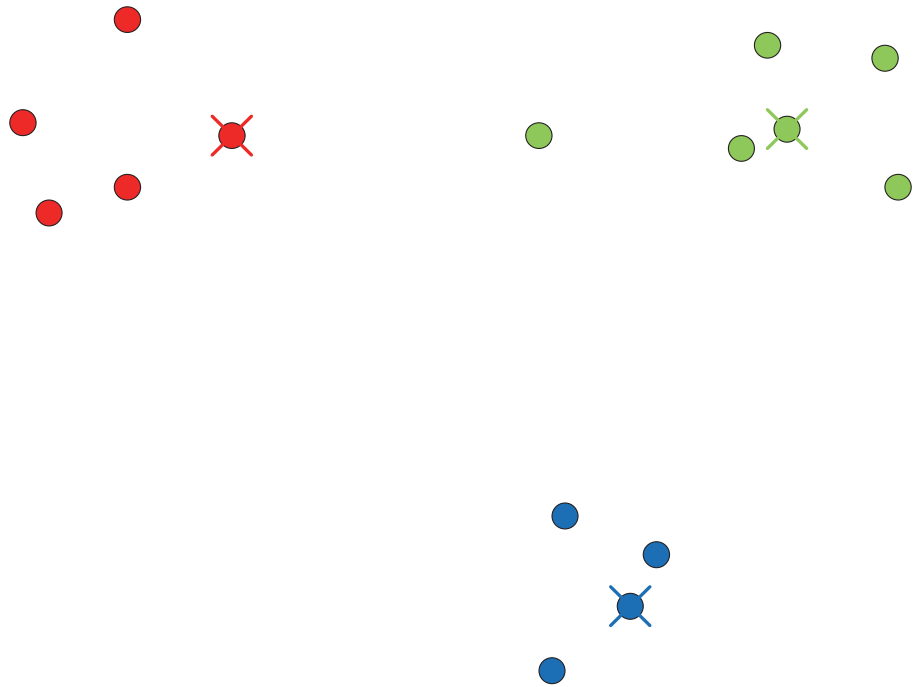
- Cost J_1^0

Iteration one - center update illustration cont.



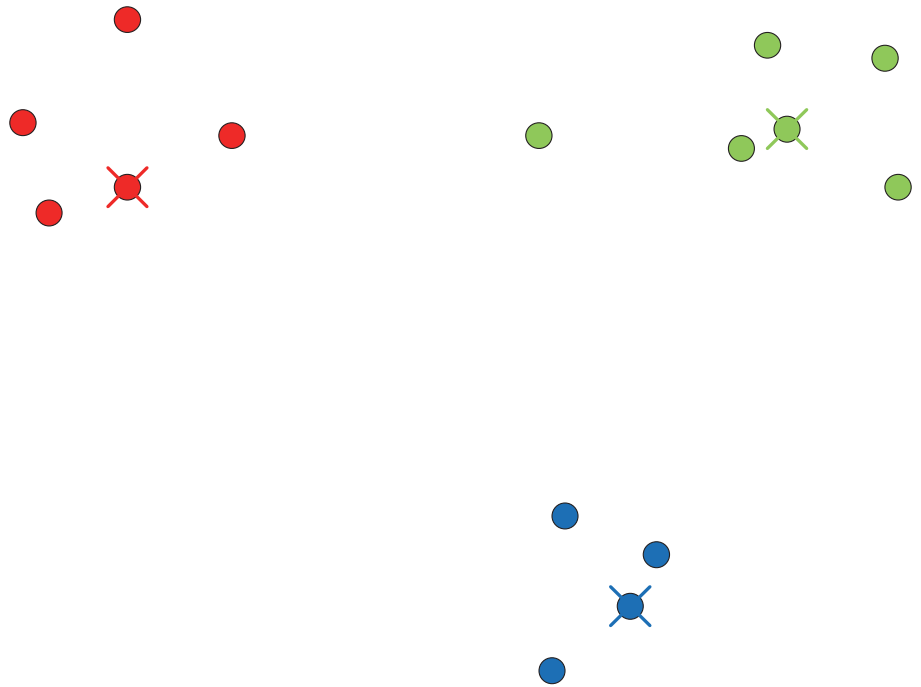
- Cost J_1^1 $(< J_1^0)$

Iteration one - cluster update illustration



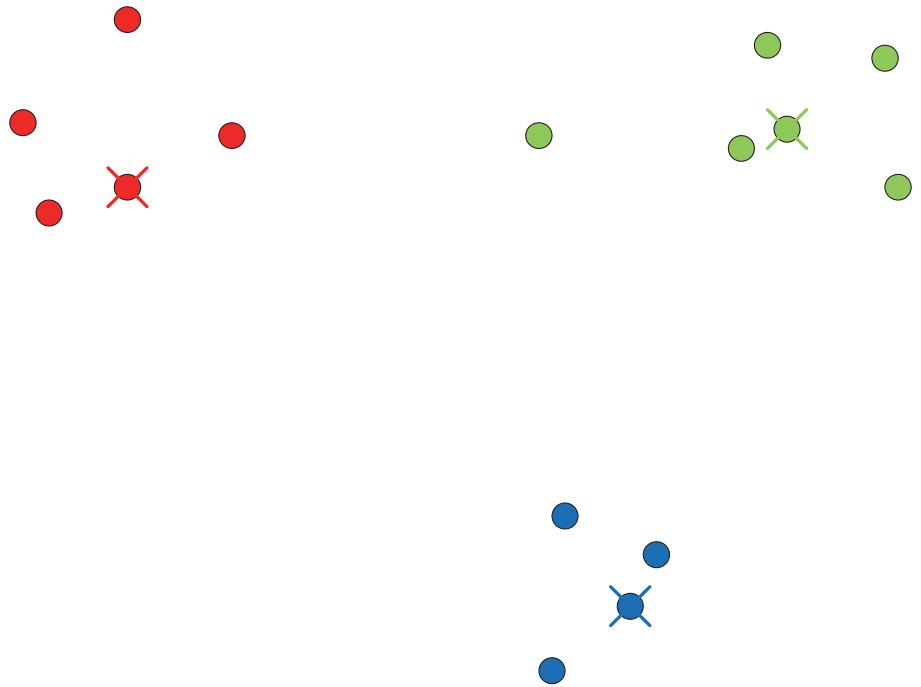
- Cost J_1^2 $(< J_1^1)$

Iteration two - center update illustration



- Cost J_1^3 $(< J_1^2)$

Iteration two - cluster update illustration



- Cost J_1^4 $(= J_1^3)$

Theoretical result - Algorithm 2

Theorem 1

Algorithm 2 converges after finite number of iterations and the error probability after T iterations is upper bounded by

$$P_e \leq 2M(K^2 + 3(T+1)(K-1)) \exp\Big(-\frac{nD_{KS}^2}{8}\Big).$$

Clustering without K

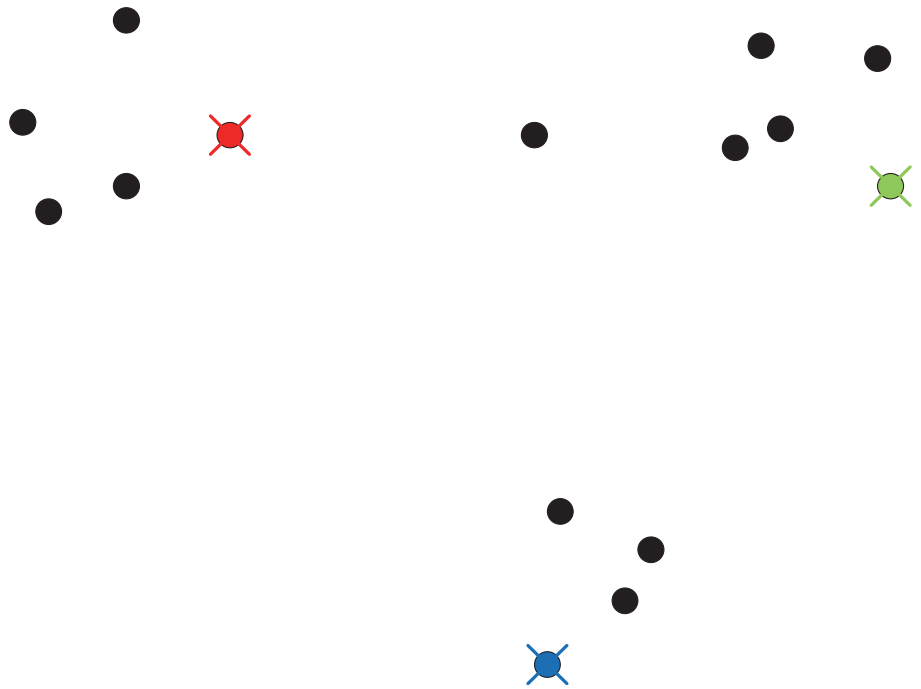
- Question: Is it possible to cluster with unknown K
  - Merge [Wang-etal 2018]
  - Split [Mora-López 2015]

Clustering without K - Merge

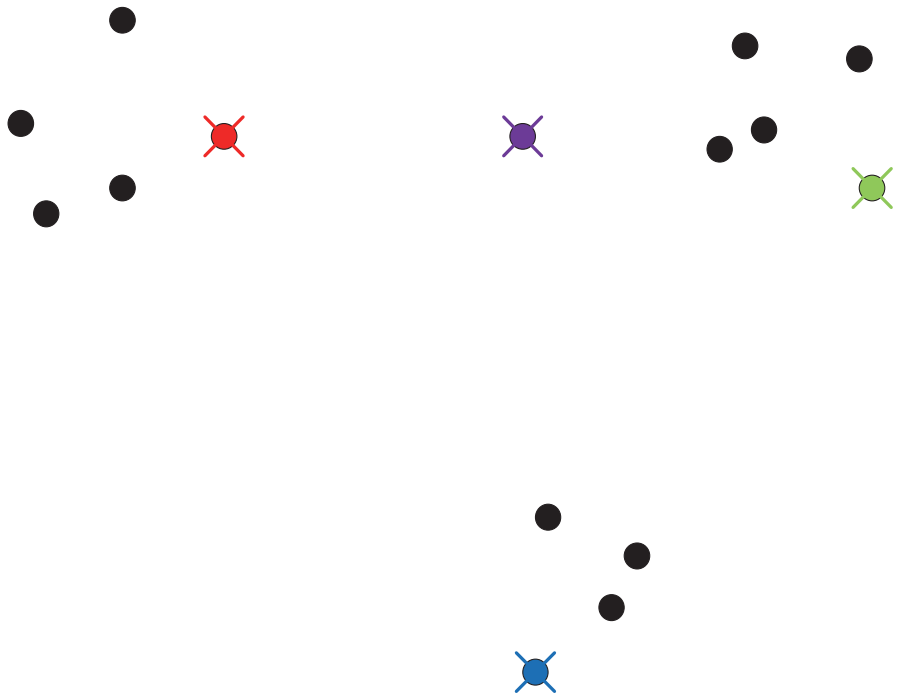
Algorithm 3 KS-based initialization with unknown K for composite distributions - merge

1: **Input**: $\{\mathbf{x}_j\}_{j=1}^{M}$.
2: $\{$Center initialization with threshold $d_{th}\}$
3: Clustering initialization specified in Algorithm 1.
4: **Output**: $\{\mathcal{C}_k\}_{k=1}^{\hat{K}}$.

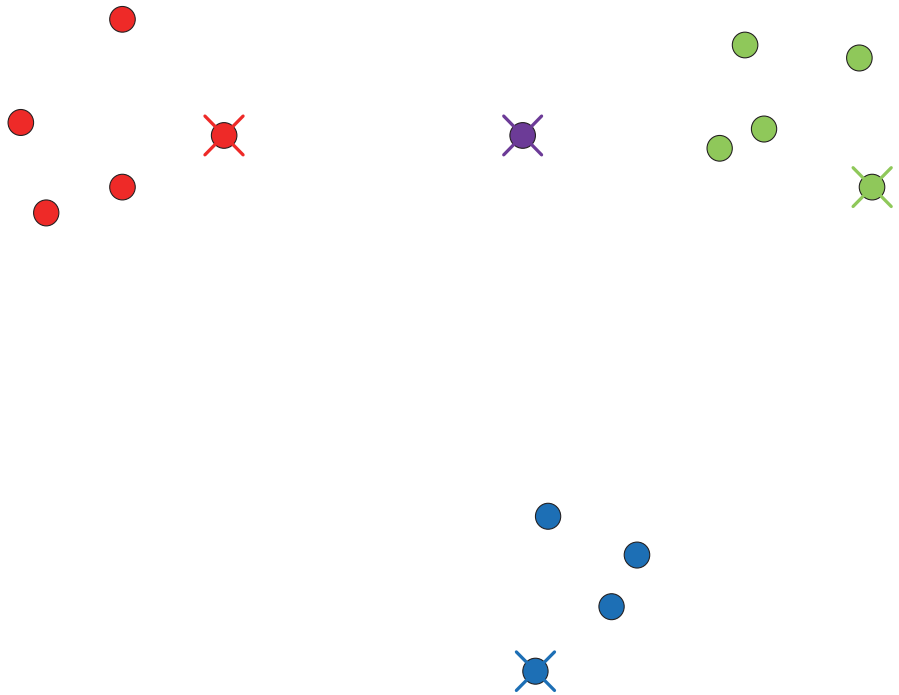
Center initialization illustration



Merge based - center initialization



Merge based - cluster initialization

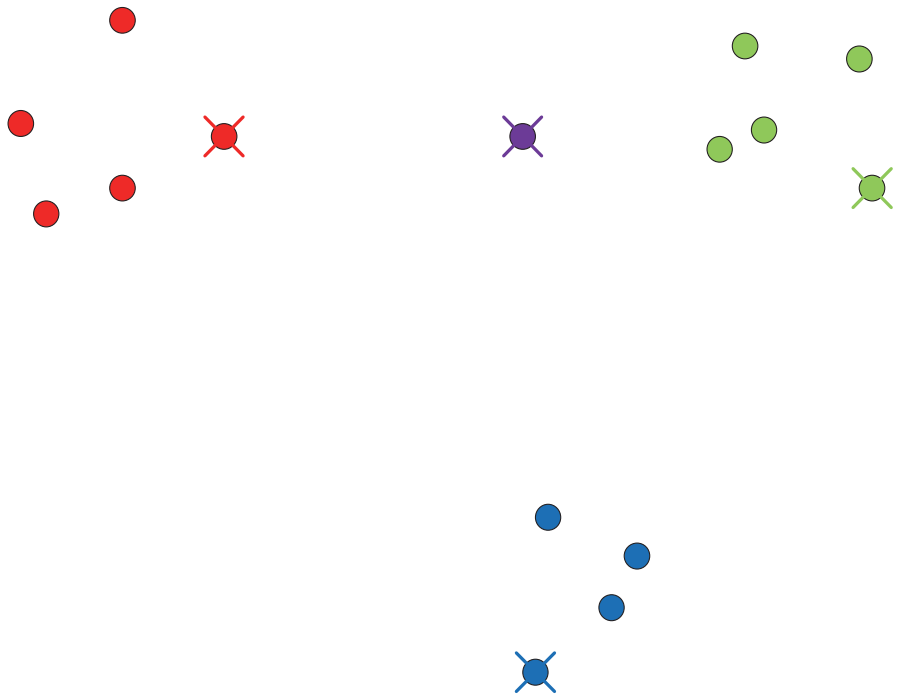


- Cost J_2^0

Algorithm 4 KS based clustering with unknown K for composite distributions - merge

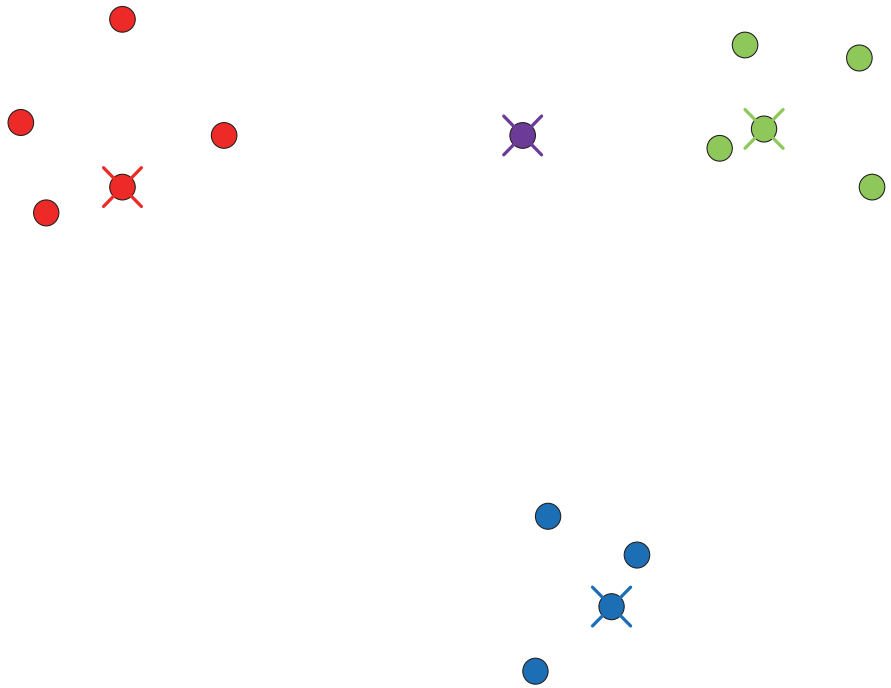
1: **Input**: $\{\mathbf{x}_j\}_{j=1}^{M}$.
2: **Initialization**: $\{\mathcal{C}_k\}_{k=1}^{\hat{K}}$ by Algorithm 3.
3: **while** the clustering result does not converge **do**
4: Center update specified in Algorithm 2.
5: {Merge Step with threshold d_{th}}
6: Cluster update specified in Algorithm 2.
7: **end while**
8: **Output**: Partition set $\{\mathcal{C}_k\}_{k=1}^{\hat{K}}$.

Merge based - center update



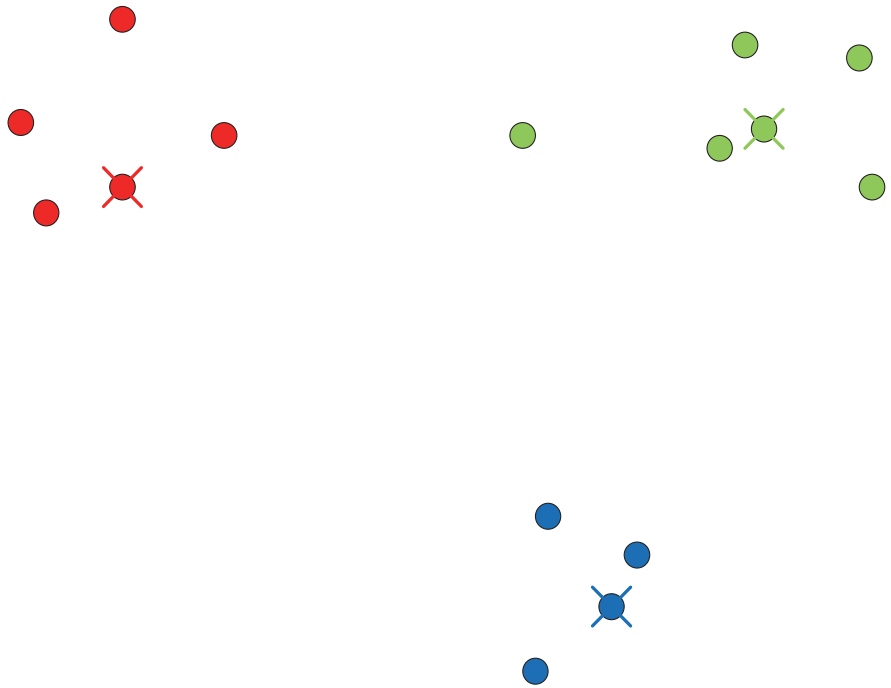
- Cost J_2^0

Merge based - center update cont



- Cost J_2^1 $(< J_2^0)$

Merge based - merge step



- Cost J_2^2

Theoretical result - Algorithm 4

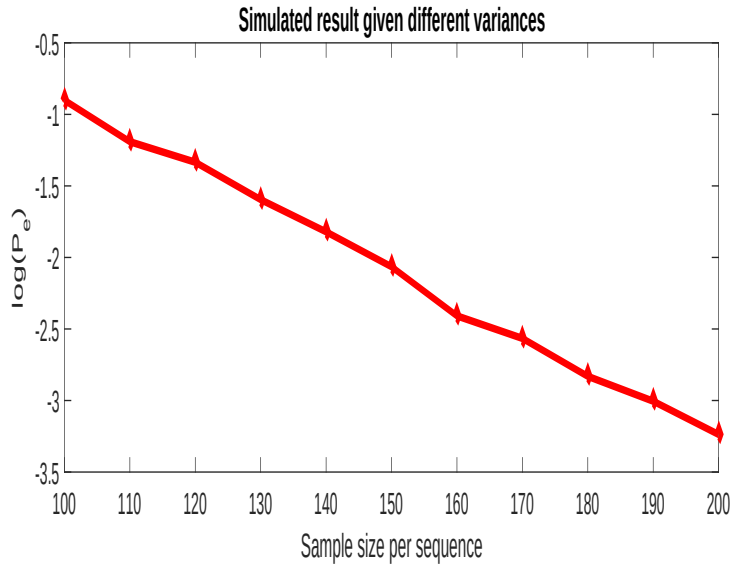
Theorem 2

Given $d_{th} = \frac{D_{KS}}{2}$, *Algorithm 3 and 4 converges after finite number of iterations and the probability of error is upper bounded by*

$$P_e \leq \left(4M^2(K+1) + 6M(K-1)(T+1) + 4TK^2\right) \exp\left(-\frac{nD_{KS}^2}{8}\right).$$

- $0 < D_{KS} \leq \min_{k \neq k'} d_{KS}(p_k, p_{k'})$

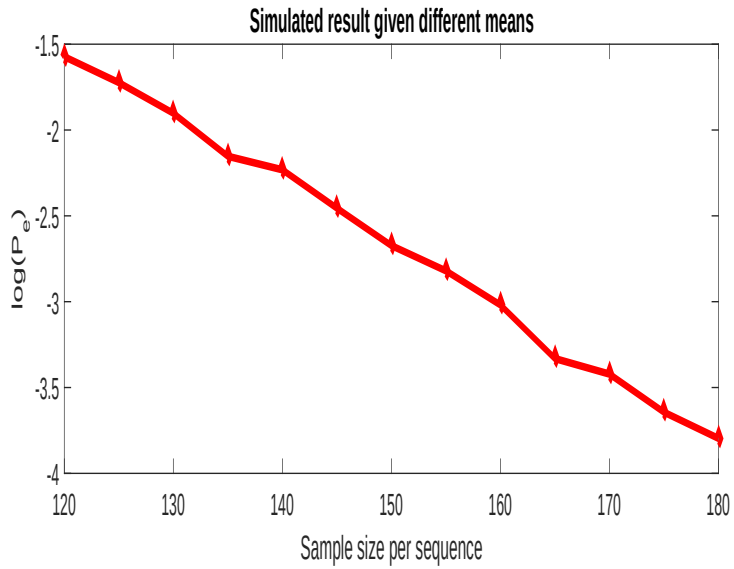
Numerical Result - Given K

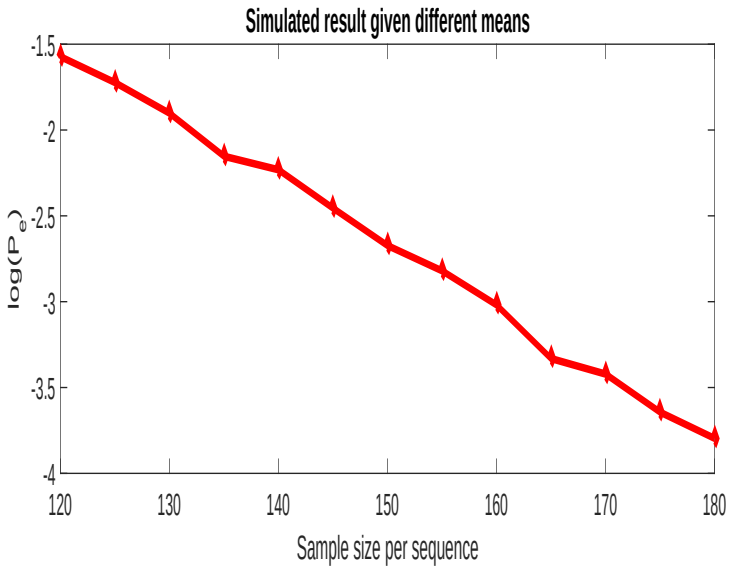


- $\mathbf{x}_j \sim \mathcal{N}(\mu, 1)$
- $\mu \in \{0, 1, 2, 3, 4\}$

- $\mathbf{x}_j \sim \mathcal{N}(0, \sigma^2)$
- $\sigma^2 \in \{1, 2, 4, 8, 16\}$

Numerical Result - Without K





- $\mathbf{x}_j \sim \mathcal{N}(\mu, 1)$
- $\mathbf{x}_j \sim \mathcal{N}(0, \sigma^2)$

Conclusion

- Clustering sequences generated by unknown continuous distributions.
  - Upper bound of P_e
  - Exponential consistency of P_e