Online Clustering with Nearly Optimal Consistency

T-H. Hubert Chan[1], Shaofeng H.-C. Jiang[2]
Tianyi Wu[2], Mengshi Zhao[1]

[1]The University of Hong Kong, [2]Peking University

April 8, 2025

Online Clustering

- **k-Means:**

Online Clustering

- **k-Means:**
  - **Input:** a data set $P \subseteq \mathbb{R}^d$ with an aspect ratio Δ.

Online Clustering

- **k-Means:**
  - **Input:** a data set $P \subseteq \mathbb{R}^d$ with an aspect ratio Δ.
  - **Goal:** find a set of k centers $C \subset \mathbb{R}^d$ to minimize the cost function

$$\mathrm{cost}(P, C) := \sum_{x \in P} (\mathrm{dist}(x, C))^2.$$

Online Clustering

- **k-Means:**
  - **Input:** a data set $P \subseteq \mathbb{R}^d$ with an aspect ratio Δ.
  - **Goal:** find a set of k centers $C \subset \mathbb{R}^d$ to minimize the cost function

$$\mathrm{cost}(P, C) := \sum_{x \in P} (\mathrm{dist}(x, C))^2.$$

- **Online k-Means:**

Online Clustering

- **k-Means:**
  - **Input:** a data set $P \subseteq \mathbb{R}^d$ with an aspect ratio Δ.
  - **Goal:** find a set of k centers $C \subset \mathbb{R}^d$ to minimize the cost function

$$\mathrm{cost}(P, C) := \sum_{x \in P} (\mathrm{dist}(x, C))^2.$$

- **Online k-Means:**
  - Data points arrive in an arbitrary order (assume points are in $[\Delta]^d$). The algorithm must decide whether and where to define a new center when a point arrives.

Online Clustering

- **k-Means:**
  - **Input:** a data set $P \subseteq \mathbb{R}^d$ with an aspect ratio Δ.
  - **Goal:** find a set of k centers $C \subset \mathbb{R}^d$ to minimize the cost function

$$\mathrm{cost}(P, C) := \sum_{x \in P} (\mathrm{dist}(x, C))^2.$$

- **Online k-Means:**
  - Data points arrive in an arbitrary order (assume points are in $[\Delta]^d$). The algorithm must decide whether and where to define a new center when a point arrives.
  - **Competitive ratio:** the ratio between the algorithm's k-Means cost and the optimal k-Means cost (with full information).

Online Clustering

- **k-Means:**
  - **Input:** a data set $P \subseteq \mathbb{R}^d$ with an aspect ratio Δ.
  - **Goal:** find a set of k centers $C \subset \mathbb{R}^d$ to minimize the cost function

$$\text{cost}(P, C) := \sum_{x \in P} (\text{dist}(x, C))^2.$$

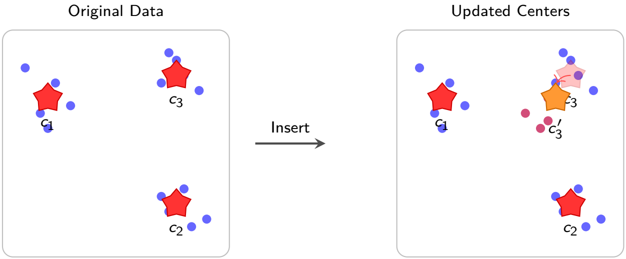
- **Online k-Means:**
  - Data points arrive in an arbitrary order (assume points are in $[\Delta]^d$). The algorithm must decide whether and where to define a new center when a point arrives.
  - **Competitive ratio:** the ratio between the algorithm's k-Means cost and the optimal k-Means cost (with full information).
  - **Lower bound**:No bounded competitive ratio [LSS16].

Online Clustering

- **Consistent online k-Means:** Recourse of decisions is allowed [LV17].

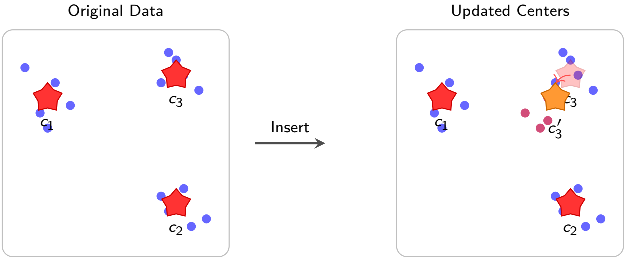
Online Clustering

- **Consistent online k-Means:** Recourse of decisions is allowed [LV17].



Online Clustering

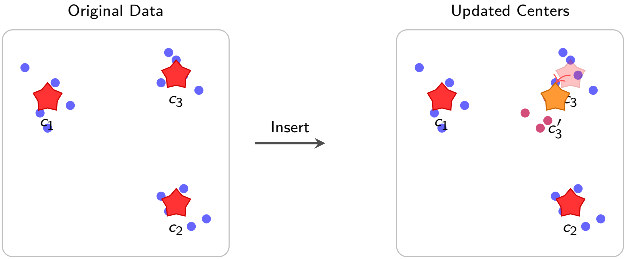
- **Consistent online k-Means:** Recourse of decisions is allowed [LV17].



- **Consistency:** $\sum_i |C_i \setminus C_{i-1}|$ where C_i is the center set after the algorithm processes the i-th input point.

Online Clustering

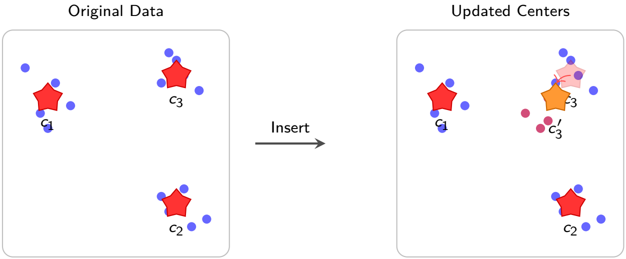
- **Consistent online k-Means:** Recourse of decisions is allowed [LV17].



  - **Consistency:** $\sum_i |C_i \setminus C_{i-1}|$ where C_i is the center set after the algorithm processes the i-th input point.
  - **Goal:** Minimizing the consistency while maintaining a bounded competitive ratio.

Online Clustering

- **Consistent online k-Means:** Recourse of decisions is allowed [LV17].



- **Consistency:** $\sum_i |C_i \setminus C_{i-1}|$ where C_i is the center set after the algorithm processes the i-th input point.
- **Goal:** Minimizing the consistency while maintaining a bounded competitive ratio.
- **Lowerbound:** any $O(1)$-competitive algorithms for k-Means must be $\Omega(k \log(n))$-consistent [LV17].

Previous Work and Limitations

- Relative works:

Algs	Ratio	Consistency	Problems
[LV17]	$O(1)$	$O(k^2 \operatorname{polylog}(n))$	k-Means and k-Median
[FLNS21]	$O(1)$	$O(k \operatorname{polylog}(n))$	k-Median

- **Gaps:**
  - If k-Means also admits $O(1)$-competitive ratio with $O(k \operatorname{polylog}(n))$-consistency?
  - Moreover, $(1+\epsilon)$-competitive ratio (might require an exponential running time)?
  - A framework, turning any offline clustering algorithm into an online algorithm with good consistency?
- **Question:** Can we close these gaps?

Main Contributions

Theorem

Given an offline α-approximate algorithm for k-Means *that runs in $T(n)$ time, there exists an $\tilde{O}_\epsilon(k)$[a]-consistent $(1+\epsilon)\alpha^2$-competitive algorithm for online* k-Means, *and the running time is $\tilde{O}_\epsilon(nk + k^3 \cdot T(\tilde{O}_\epsilon(k)))$.*

[a] $\tilde{O}_\epsilon(k)$ hides poly$(\epsilon^{-1} d \log(n))$ factor.

Main Contributions

Theorem

Given an offline α-approximate algorithm for k-Means *that runs in $T(n)$ time, there exists an $\tilde{O}_\epsilon(k)$[a]-consistent $(1+\epsilon)\alpha^2$-competitive algorithm for online* k-Means, *and the running time is $\tilde{O}_\epsilon(nk + k^3 \cdot T(\tilde{O}_\epsilon(k)))$.*

[a] $\tilde{O}_\epsilon(k)$ hides $\mathrm{poly}(\epsilon^{-1} d \log(n))$ factor.

- this result more generally works for (k, z)-Clustering (which particularly contains k-Median), in addition to k-Means.

Main Contributions

Theorem

Given an offline α-approximate algorithm for k-Means *that runs in $T(n)$ time, there exists an $\tilde{O}_\epsilon(k)$[a]-consistent $(1+\epsilon)\alpha^2$-competitive algorithm for online* k-Means, *and the running time is $\tilde{O}_\epsilon(nk + k^3 \cdot T(\tilde{O}_\epsilon(k)))$.*

[a] $\tilde{O}_\epsilon(k)$ hides $\mathrm{poly}(\epsilon^{-1} d \log(n))$ factor.

- this result more generally works for (k, z)-Clustering (which particularly contains k-Median), in addition to k-Means.
- plugging in the brute-force exact offline algorithm leads to a $(1+\epsilon)$-competitive $O(k \operatorname{polylog} n)$-consistent algorithm for k-Means.

Main Contributions

Theorem

Given an offline α-approximate algorithm for k-Means *that runs in $T(n)$ time, there exists an $\tilde{O}_\epsilon(k)$[a]-consistent $(1+\epsilon)\alpha^2$-competitive algorithm for online* k-Means, *and the running time is $\tilde{O}_\epsilon(nk + k^3 \cdot T(\tilde{O}_\epsilon(k)))$.*

[a] $\tilde{O}_\epsilon(k)$ hides $\mathrm{poly}(\epsilon^{-1} d \log(n))$ factor.

- this result more generally works for (k, z)-Clustering (which particularly contains k-Median), in addition to k-Means.
- plugging in the brute-force exact offline algorithm leads to a $(1+\epsilon)$-competitive $O(k \,\mathrm{polylog}\, n)$-consistent algorithm for k-Means.

Algorithm	**Ratio**	**Consistency**
[LV17](k-Median, k-Means)	$O(1)$	$\tilde{O}(k^2)$
[FLNS21](k-Median)	$O(1)$	$\tilde{O}(k)$
Our work (k-Median, k-Means**)**	$1+\epsilon$	$\tilde{O}_\epsilon(k)$

Key Ideas

- Consistent Coreset Construction: reduce the number of points to be considered: from n points to $\tilde{O}_\epsilon(k)$ weighted points.

coreset

A weighted set $S \subseteq \mathbb{R}^d$ such that $\forall C \subseteq \mathbb{R}^d, |C| = k$,
$\operatorname{cost}(S, C) \in (1 + \epsilon) \operatorname{cost}(P, C)$.

Key Ideas

- Consistent Coreset Construction: reduce the number of points to be considered: from n points to $\tilde{O}_{\epsilon}(k)$ weighted points.

coreset

A weighted set $S \subseteq \mathbb{R}^d$ such that $\forall C \subseteq \mathbb{R}^d, |C| = k$,
$\mathrm{cost}(S, C) \in (1 + \epsilon)\,\mathrm{cost}(P, C)$.

- Algorithms For Bounded Input: achieving $\tilde{O}_{\epsilon}(1)$-amortized consistency.

Key Ideas

- Consistent Coreset Construction: reduce the number of points to be considered: from n points to $\tilde{O}_\epsilon(k)$ weighted points.

coreset

A weighted set $S \subseteq \mathbb{R}^d$ such that $\forall C \subseteq \mathbb{R}^d, |C| = k$,
$\mathrm{cost}(S, C) \in (1+\epsilon)\,\mathrm{cost}(P, C)$.

- Algorithms For Bounded Input: achieving $\tilde{O}_\epsilon(1)$-amortized consistency.
- $\tilde{O}_\epsilon(1)$-amortized consistency + $\tilde{O}_\epsilon(k)$ points $\rightarrow$ $\tilde{O}_\epsilon(k)$-consistency.

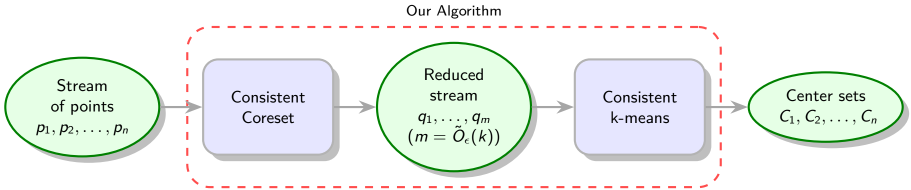
Key Ideas

- Consistent Coreset Construction: reduce the number of points to be considered: from n points to $\tilde{O}_\epsilon(k)$ weighted points.

coreset

A weighted set $S \subseteq \mathbb{R}^d$ such that $\forall C \subseteq \mathbb{R}^d, |C| = k$, $\operatorname{cost}(S, C) \in (1+\epsilon)\operatorname{cost}(P, C)$.

- Algorithms For Bounded Input: achieving $\tilde{O}_\epsilon(1)$-amortized consistency.
- $\tilde{O}_\epsilon(1)$-amortized consistency + $\tilde{O}_\epsilon(k)$ points $\rightarrow$ $\tilde{O}_\epsilon(k)$-consistency.



Experiment

Datasets

Dataset	Dimension	Size
SKIN [BD09]	3	245057
SHUTTLE [Cat]	7	58000
COVERTYPE [Bla98]	54	581012

Algorithms:

- Baseline1 (LV17): Algorithm in [LV17].
- Baseline2 (naive): Directly running k-Means++ on the consistent coreset instead of running any additional algorithm.
- Our algorithm (ours-faithful): plugging k-Means++ to our framework.
- A heuristic implementation of our algorithm (ours-heuristic): when a new point is added, it replaces only one existing center if doing so maximizes cost reduction.

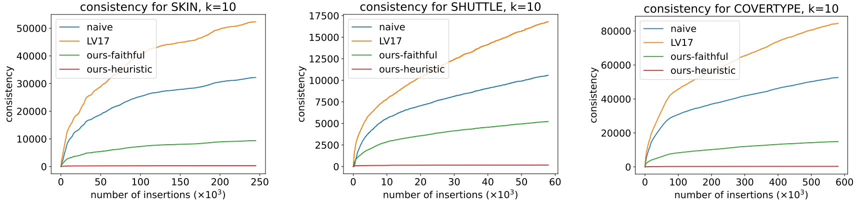


Figure: The consistency curve over the insertions of points, for all datasets and $k = 10$.

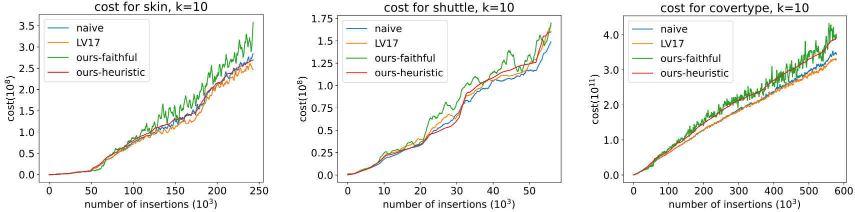


Figure: The cost curve over the insertions of points, for all datasets and $k = 10$. We plot the curve after applying a moving average with a window size equal to 1% of the dataset size.

Thank you!

Rajen Bhatt and Abhinav Dhall.
Skin Segmentation.
UCI Machine Learning Repository, 2009.
DOI: https://doi.org/10.24432/C5T30C.

Jock Blackard.
Covertype.
UCI Machine Learning Repository, 1998.
DOI: https://doi.org/10.24432/C50K5N.

Jason Catlett.
Statlog (Shuttle).
UCI Machine Learning Repository.
DOI: https://doi.org/10.24432/C5WS31.

Hendrik Fichtenberger, Silvio Lattanzi, Ashkan Norouzi-Fard, and Ola Svensson.
Consistent k-clustering for general metrics.
In *SODA*, pages 2660–2678. SIAM, 2021.

Edo Liberty, Ram Sriharsha, and Maxim Sviridenko.

An algorithm for online k-means clustering.
In *ALENEX*, pages 81–89. SIAM, 2016.

Silvio Lattanzi and Sergei Vassilvitskii.
Consistent k-clustering.
In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 1975–1984. PMLR, 2017.