Appendix A: supplementary materials to Section 3.1

In this part of the Appendix, we provide details on the construction of our framework that are not included in Section 3.1 due to space constraints.

Handling degenerate and boundary points

One problem with $k$-means is it may produce degenerate solutions: if the solution $C'$ has $k$ centroids, it is possible that data points are mapped to only $k' < k$ centroids. To handle degenerate cases, starting with $|C^0| = k$, we consider an enlarged clustering space $\{A\}_k$, which is the union of all $k'$-clusterings with $1 \leq k' \leq k$. We use the pre-image $v^{-1}(A) \in \{C\}$ to denote the non-boundary points $C$ such that $v(C) = A$, i.e., these are the set of non-boundary points in the equivalence class induced by clustering $A$. To include boundary points as well, we devise the operator $C(l)$ as the “closure” of an equivalence class $v^{-1}(A)$, which includes all boundary points $C'$ such that $A \in V(C') \cap X$.

Using the above two extensions, we give the robust definition of stationary clusterings and stationary points, which we use in our analysis.

Definition 7 (Stationary clusterings). We call $A^*$ a stationary clustering of $X$, if $m(A^*) \in Cl(v^{-1}(A^*))$. We let $\{A\}_k \subseteq \{A\}_k$ denote the set of all stationary clusterings of $X$ with number of clusters $k' \in [k]$.

For each $A^*$, we define a matching centroidal solution $C^*$. Definition 8 (Stationary points). For a stationary clustering $A^*$ with $k'$ centroids, we define $C^* = \{c_r^*, r \in [k]\}$ to be a stationary point corresponding to $A^*$, so that $\forall A_r^* \in A^*$, $c_r^* := m(A_r^*)$. We let $\{C^*\}_k$ denote the corresponding set of all stationary points of $X$ with $k'$ in $[k]$.

With the robust definitions, Figure 3 provides a visualization of batch $k$-means walking on $\{C\}$ (and $\{A\}_k$) as an iterative mapping $m \circ v$ ($v \circ m$, resp.). In $\{C\}$, it jumps from one equivalence class to another until it stays in the same equivalence class in two consecutive iterations.

Now we extend $\Delta(\cdot, \cdot)$ to include the degenerate cases. Fix a clustering $A$ with its induced $k$ centroids $C := m(A)$, and another set of $k'$-centroids $C' (k' \geq k)$ with its induced clustering $A'$, if $|A'| = |A| = k$ (this means if $k' > k$, then $C'$ has at least one degenerate centroid), then we can pair the subset of non-degenerate $k$ centroids in $C'$ with those in $C$, and ignore the degenerate centroids. Under this condition, we can extend Definition 2 to include degenerate solutions as well, provided $C = m(A)$ for some clustering $A$, which is always satisfied in our subsequent analysis.

A sufficient condition for the local convergence of batch $k$-means

We show batch $k$-means algorithm has geometric convergence in the local neighborhood of a stable stationary point in the solution space.

Proof of Lemma 2. Without loss of generality, we let $\tau(r) = r, \forall r \in [k]$. Let $\rho_{out} := \frac{|\Delta x_r(A_r \cap A'_r)|}{n_r^2}$, and $\rho_{in} := \frac{|\Delta x_r(A_r \cap A'_r)|}{n_r}$; let $\rho_{max} := \max_r \frac{\rho_{out} + \rho_{in}}{n_r}$.

Clearly, $(\rho_{out} + \rho_{in}) \leq \rho_{max}$, by our definition. Now, similar to [19], we can get $\|m(A_r) - c_r^*\| \leq \frac{1 - \rho_{out}}{1 - \rho_{out} + \rho_{in}} \|m(A_r \cap A'_r) - c_r^*\| + \frac{\rho_{in}}{1 - \rho_{out} + \rho_{in}} \|m(A_r \cap A'_r) - c_r^*\| + \frac{\rho_{in}}{1 - \rho_{out} + \rho_{in}} \|m(A_r \cap A'_r) - c_r^*\|
And as in [19], we get $(1 - \rho_{out}) \|m(A_r \cap A'_r) - c_r^*\| < \frac{1}{2} \|m(A_r - c_r^*)\|$. Now we bound the second term: by Cauchy-Schwarz inequality, $\|m(A_r - c_r^*)\| \leq \sum_{s \neq r} n_r \|m(A_r) - c_r^*\| \leq \phi(C') + \sum_r n_r \|m(A_r) - c_r^*\| \leq \phi(C') + \sum_r \|m(A_r) - c_r^*\|^2 \leq \phi(C) + \sum_r \|m(A_r) - c_r^*\|^2$. By Lemma B.3, $\sum_r \sum_{i \neq r} n_r \|m(A_r \cap A'_i) - c_i^*\|^2$ can be upper bounded by $\phi(C') + \sum_r n_r \|m(A_r) - c_r^*\|^2 \leq \phi(C) + \sum_r \|m(A_r) - c_r^*\|^2 \leq \phi(C') + \sum_r \|m(A_r) - c_r^*\|^2$. Substituting this into the previous inequality, we have $\sum_r n_r \phi(C' + \sum_r n_r \|m(A_r) - c_r^*\|^2 \leq \phi(C') + \sum_r \|m(A_r) - c_r^*\|^2 \leq \phi(C) + \phi'(C')$. Thus, $\sum_r n_r \|m(A_r) - c_r^*\|^2 \leq \frac{1}{b} \sum_r n_r \|m(A_r) - c_r^*\|^2 \leq \frac{1}{b} \sum_r n_r \|m(A_r) - c_r^*\|^2 \leq \frac{1}{b} \sum_r n_r \|m(A_r) - c_r^*\|^2 \leq \frac{1}{b} \sum_r n_r \|m(A_r) - c_r^*\|^2 < 1$. Therefore, $\sum_r n_r \|m(A_r) - c_r^*\|^2 \leq \frac{1}{b} \sum_r n_r \|m(A_r) - c_r^*\|^2 \leq \frac{1}{b} \sum_r n_r \|m(A_r) - c_r^*\|^2 < 1$. Hence, $\phi(C') \leq \phi(C)$ (equality holds if $C$ is a stationary point).}

Lemma 3. Fix any target clustering $C^*$, and another clustering $C$ with a matching $\pi : [k] \rightarrow [k]$. Let $C' := \{A^*_r\}$ }
\{m(A_r), r \in [k]\}. Then
\[
\sum \sum_{r \neq r'} \sum_{x \in A_r \cap A_r'} \|x - c^*_r\|^2 \\
\leq \phi(C') - \sum_{r} \phi(c^*_r; A_{p(r)} \cap A_r') + \sum_{r} \eta_r \|m(A_r) - c^*_r\|^2
\]

Proof. Without loss of generality, we let \(\pi(r) = r\).
\[
\phi(C') - \phi(C^*) = \sum_{r \neq r'} \sum_{x \in A_r} \|x - c^*_r\|^2 - \sum_{r} \sum_{x \in A_r} \|x - c^*_r\|^2 + \sum_{r \neq r'} \sum_{x \in A_r} \|x - c^*_r\|^2
\]

So \(\sum_{r} \sum_{x \in A_r} \|x - c^*_r\|^2 - \sum_{r} \sum_{x \in A_r} \|x - c^*_r\|^2 = \phi(C') - \phi(C^*) - \sum_{r} \sum_{x \in A_r} \|x - m(A_r)\|^2 + \sum_{r} \sum_{x \in A_r} \|x - c^*_r\|^2 \leq \phi(C) - \phi(C^*) + \sum_{r} \eta_r \|m(A_r) - c^*_r\|^2\).

Now, we claim \(\sum_{r} \sum_{x \in A_r} \|x - c^*_r\|^2 - \sum_{r} \sum_{x \in A_r} \|x - c^*_r\|^2 = \sum_{r} \sum_{x \in A_r} \|x - c^*_r\|^2 + \sum_{r} \sum_{x \in A_r} \|x - c^*_r\|^2 \leq \phi(C) - \phi(C^*) + \sum_{r} \eta_r \|m(A_r) - c^*_r\|^2 + \sum_{r} \sum_{x \in A_r} \|x - c^*_r\|^2\).

This is because we can enumerate \(x\) using clustering \(A_r\) for each \(x \in A_r\), either \(x \in A_r \cap A_r'\), then \(\|x - c^*_r\|^2 - \|x - c^*_r\|^2 = 0\), or \(x \in A_r \cup A_r'\) for some \(s \neq r\), which means the difference is \(\|x - c^*_r\|^2 - \|x - c^*_r\|^2\).

(And this term is positive by optimality of clustering \(A_r\), fixing \(\{c^*_r\}\)).

Therefore, \(V(C') \cap X = A\), i.e., \(C' \in v^{-1}(A)\).

8 Appendix B: Local Lipschitzness and clusterability

Lemma 4. The following are equivalent

1. \(C\) is a boundary point
2. \(V(C)\) has a zero margin with respect to \(X\)
3. \(|V(C) \cap X| > 1\), i.e., the clustering determined by \(V(C)\) is not unique.

Proof of Lemma 4 “1 \(\implies\) 2’’ obviously holds since \(\|x - c^*_r\| = \|x - c^*_r\|\) if and only if \(\|x - c^*_r\| = \|x - c^*_r\|\).

“2 \(\implies\) 3”: let \(A \in V(C) \cap X\) be the clustering achieving the zero margin, and consider \(x \in A \cup A\), s.t. \(\|x - c^*_r\| \neq \|x - c^*_r\|\); without loss of generality, assume \(x \in A\), according to clustering \(A\), and define \(A'\) to be the same clustering as \(A\) for all points in \(X\) but \(x\), where it assigns \(x\) to \(A\). Then \(A' \in V(C) \cap X\) and \(|V(C) \cap X| \geq 2\). “3 \(\implies\) 1”: Suppose otherwise. Then every point \(x\) has a unique center that minimizes its distance to it, which means the clustering determined by \(V(C)\) \(\cap X\) is unique. A contradiction.

Lemma 5. If \(C^* \in \{C^*\}[k]\), then \(C^* = m(A^*)\), where \(A^* \in \{A^*\}\) and \(A^* = v(C^*)\).

Proof. By definition of stationary points, \(C^* = m \circ v(C^*)\). Let \(A = v(C^*)\), then \(m(A) = C^*\) and \(v(m(A)) = v(C^*) = A\). Thus \(A \in \{A^*\}\) by definition of a stationary clustering.

Lemma 6. Fix a clustering \(A = \{A_1, \ldots, A_k\}\), and let \(C \in v^{-1}(A)\). Then \(\exists \delta > 0\) such that the following statement holds:

\[\Delta(C', C) < \delta \implies C' \in v^{-1}(A)\]

Proof. Since \(C\) is not a boundary point, \(\forall x \in A_r, r \in [k]\),
\[\|x - c_r\| < \|x - c_s\|, \forall s \neq r\]
So we can choose \(\delta > 0\) s.t. \(\forall x \in A_r, \forall r \in [k], s \neq r\),
\[\|x - c_r\| < \|x - c_s\| - 2\delta\]
Let \(\pi^*\) be a permutation such that \(\Delta(C', C)\) is defined. We have \(x \in A_r, r \in [k], s \neq r\),
\[\|x - c^*_{r(r)} - c^*_{s(s)}\| < \delta\]
where the second inequality is by the fact that
\[\max_r \|c^*_{r(r)} - c^*_{s(s)}\| < \delta\]
Therefore, \(V(C') \cap X = A\), i.e., \(C' \in v^{-1}(A)\).

Lemma 7. Suppose \(\forall C^* \in \{C^*\}[k]\), \(C^*\) is not a boundary point (i.e., suppose Assumption \(A\) holds). Let \(C = m(A^*) \notin \{C^*\}[k]\) for some \(A^* \in \{A\}\) and let \(C^* \in Cl(v^{-1}(A^*))\), then \(\exists \delta > 0\) s.t. \(\Delta(C', C) \geq \delta\).

Proof. We prove the lemma by contradiction: suppose \(\forall \delta > 0\), \(\exists C'\) s.t. \(C' \in Cl(v^{-1}(A^*))\) and \(\Delta(C', C) < \delta\).

First, we claim that for \(\delta\) sufficiently small, \(C\) must be a boundary point: suppose otherwise, then by Lemma 6 \(v(C') = v(C) = A^*\), contradicting the fact that \(C \notin \{C^*\}[k]\). Let \(A \in V(C) \cap X\). Since \(C\) is a boundary point, \(\exists r, s\) and \(x \in A_r \cup A_s\), s.t.
\[\|x - c_r\| = \|x - c_s\|\]
Now, we choose \(\delta > 0\) to be sufficiently small so that for any \(A^* \in V(C^*) \cap X\), clustering \(A^*\) only differs from \(A\) on the assignment of these points sitting on the bisection. This implies \(C \in Cl(v^{-1}(A))\), which implies \(C\) is a boundary stationary point, a contradiction.

Lemma 8. If \(\forall C^* \in \{C^*\}[k]\), \(C^*\) is a non-boundary stationary point, that is, \(C^* := m(A^*) \in v^{-1}(A^*)\). Then \(\exists \min r > 0\) such that \(\forall C^* \in \{C^*\}[k]\), \(C^*\) is a \((r_{\min}, 0)\)-stable stationary point.

Proof. Fix any \(r\) in the range of \([k]\) (we abuse the notation with the same \(k\) here). For any \(C\) such that \(\Delta(C, C^*)\) exists (i.e., \(|C| = k \geq k = \{C^*\} |\)), we first show \(\exists \delta > 0\), such that the following statement holds:
\[\Delta(C, C^*) > \delta \implies C \in v^{-1}(A^*)\]
Since \(C^*\) is a non-boundary point, there is a permutation \(\pi_o\) of \([k]\) such that \(\forall x \in A_r, \forall r \in [k]\) and \(\forall s \neq r\),
\[\|x - c^*_{\pi_o(s)}\| < \|x - c^*_{\pi_o(s)}\|\]
We choose $r^* > 0$ so that $\forall x \in A_r$, $\forall r \in [k]$, $\forall s \neq r,$
\[\|x - c^*_{\pi(r)}\| \leq 2\sqrt{r^*}\phi^*; \; \forall r \in [k], \; s \neq r\]
with equality holds for at least one triple of $(x, r, s)$. Let $\pi^*$ be a permutation satisfying
\[\pi^* = \arg \min_{\pi} \sum_{r \in [k]} n^*_r \|c_{\pi(r)} - c^*_r\|^2\]
Let $\pi' := \pi^* \circ \pi_0$. We have $\forall (x, r, s),$
\[\|x - c_{\pi'(r)}\| - \|x - c_{\pi'(s)}\| \geq \|x - c_{\pi_0(r)}\| - 2\sqrt{r^*}\phi^* > 0\]
where the second inequality is by the fact that
\[
\max_r \|c_{\pi(r)} - c^*_r\|^2 \leq \Delta(C, C') \leq r^*\phi^* \Rightarrow \max_r \|c_{\pi(r)} - c^*_r\| < \sqrt{r^*}\phi^*
\]
Since $\pi'$ is the composition of two permutations of $[k]$, it is also a permutation of $[k]$, and $\forall r, s \neq r,$ $\|x - c_{\pi'(r)}\| < \|x - c_{\pi'(s)}\|$, so $C \in v^{-1}(r^*)$. Since by our definition, $r^*$ is unique for each $C$. Since $(C^*_{[k]})_{[k]}$ is finite, taking the minimum over all such $r^*$, i.e., $r_{\min} := \min_{C^* \in (C^*_{[k]})_{[k]}} r^*$ completes the proof.

The following is a restatement of Lemma 2 which is robust to degeneracy and boundary points.

**Lemma 9 (Restatement of Lemma 2).** If $X$ is a general dataset, then $3r_{\min} > 0$ s.t.

1. $\forall C^* \in \{C^*\}_{[k]}, \; C^* \text{ is a } (r_{\min}, 0)\text{-stable stationary point.}$

2. Let $m(A') \notin \{C^*\}_{[k]}$ for some $A' \in \{A\}$ and let $C' \in C l(v^{-1}(A'))$, then $\Delta(C', m(A)) \geq r_{\min}\phi(m(A)).$

**Proof.** By Lemma 8, $3r_{\min} > 0$ s.t. $\forall C^*, \; C^* \text{ is } r_{\min}\text{-stable.}$ Furthermore, by Lemma 2, $3r_{\min} > 0$ s.t. $\forall C^*, \Delta(C', m(A)) \geq r_{\min}\phi(m(A)).$ Let $r_{\min} := \min\{r_{\min}, r_{\min}\}$ completes the proof.

**Proof of Proposition**

For all $r \in [k],\n\[n^*_r \|c_r - c^*_r\|^2 \leq \Delta(C, C^*) \leq b\phi^*
\]
so $\|c_r - c^*_r\| \leq \sqrt{\frac{b\phi^*}{n^*_r}}$. Then for all $r \neq s,$
\[
\|c_r - c^*_r\| + \|c_s - c^*_s\| \leq \sqrt{\frac{b\phi^*}{n^*_r}} + \sqrt{\frac{b\phi^*}{n^*_s}}
\]
\[
= \sqrt{\frac{b}{f}} \sqrt{\phi} \sqrt{\frac{1}{n^*_r} + \frac{1}{n^*_s}} \leq \sqrt{\frac{b}{f}} \Delta_{rs} \leq \frac{1}{\sqrt{\Delta_{rs}}}
\]
where the second inequality is by (B), and the last inequality by our assumption on $b$. Thus, we may apply Lemma 17 to get $\frac{\alpha - \Delta_{rs}}{n^*_r} \leq \frac{1}{n^*_r}$ for all $r$, proving the first statement. Now by Lemma 18, $\phi(C) \leq (b + 1)\phi^*$, so
\[
\frac{ab}{5ab + 4(1 + 2\alpha/c)} \geq \frac{ab}{5ab + 4(2 + b)}
\]
\[
\geq \frac{5\alpha^2/16 + 4(2 + f^2/16^2)}{\beta f^3(\alpha)} \geq \frac{|A_\Delta|}{n^*_r}
\]
where the third inequality holds since $f \geq \max\{64^2, \frac{5\alpha^2 + 1}{16^3}\}$ by (B). This proves the second statement since $C^*$ is then $(f^2, \alpha)$-stable by Definition 1.

**9 Appendix C: Proof of Theorem 3.**

**Theorem 3.** Fix any $0 < \delta < \frac{1}{2}$. Suppose $C^*$ is $(b, \alpha)$-stable. If we run Algorithm 1 with parameters satisfying
\[
m > \frac{\ln(1 - \sqrt{\alpha})}{\ln(1 - 1/5p_{\min})}
\]
\[
c' > \beta \frac{\sqrt[2]{2(1 - \sqrt{\alpha}) - (1 - 5p_{\min})}}{2} \text{ with } \beta \geq 2
\]
\[
t_0 \geq 768(c')^2(1 + b_o)^2 n_o \ln n_o \frac{1}{\delta}
\]
**Then, if at some iteration $i$, $C^* \leq \frac{1}{2} b_o \phi^*$, we have $\forall t > i,$**
\[
Pr(\Omega_t) \geq 1 - \delta \text{ and }
\]
\[
E[\Delta_t^i] \leq \frac{(c')^2 B(t_o + i + 1)^\beta \Delta_t^i}{(t_o + i + 1)^{\beta + 1} (t_o + t + 1)}
\]
where $B := 4(b_o + 1)n_o\phi^*$.

**9.1 Proofs leading to Theorem 3**

In the subsequent analysis, we let
\[
\beta' := 2c' \min_{r} p_r(m) (1 - \max_{r} p_r(m) \sqrt{\alpha})
\]
so
\[
\beta' = 2c' \min_{r} p_r(m) - \sqrt{\alpha} \max_{r} p_r(m)
\]
The noise terms appearing in our analysis are:
\[
E[\sum_{r \in A^*_{t+1}} \|x - c_{\hat{r}^i} - c_r\|^2 + \phi^* | F_i]
\]
\[
\sum_{r} n^*_r (c^*_r - c_{\hat{r}^i} - c_r - E[c_{\hat{r}^i} | F_{i-1}])
\]
\[
\sum_{r} n^*_r (c^*_r - c_r)^2
\]
In the analysis of this section, we use $E_{t}[\cdot]$ as a shorthand notation for $E[\cdot|\Omega_{t}]$, where $\Omega_{t}$ is as defined in the main paper. Let $F_{t}$ denote the natural filtration of the stochastic process $C_{0}, C_{1}, \ldots$, up to $t$.

The main idea of the proof is to show that with proper choice of the algorithm’s parameters $m, c', and \ell_{o}$, the following holds at every step $t$:

- $\beta^{t} \geq 2 |\Omega_{t}|$
- Noise terms $[11]$ and $[12]$ are upper bounded by a function of $\phi_{t}^{\alpha}|\Omega_{t}|$
- $Pr(\Omega_{t} \setminus \Omega_{t+1})$ is negligible $|\Omega_{t}, \beta^{t} \geq 2$, bounded noise
- $E_{t}[\Delta^{*}|F_{t-1}] \leq (1 - \frac{\beta^{t}}{(t + \alpha)})\Delta^{*} + c' |\Omega_{t}$

where $c'$, the noise term, decreases of order $O(\frac{1}{t})$.

**Lemma 10.** Suppose $C^{*}$ is $(b_{o}, \alpha)$-stable. If

$$m > \frac{\ln(1 - \sqrt{\delta})}{\ln(1 - \frac{1}{2}p_{min}^{*})}$$

and

$$c' > \frac{\beta}{2[1 - \sqrt{\delta} - (1 - \frac{1}{2}p_{min}^{*})m]}$$

Then conditioning on $\Omega_{t}$, we have $\beta^{t} \geq \beta$.

**Proof.** Let’s first consider $p_{\ell}^{r}(1) = \frac{n_{r}^{\ell-1}}{n}$. Conditioning on $\Omega_{t}$, using the fact that $C^{*}$ is $(b_{o}, \alpha)$-stable, we have

$$n_{r}^{\ell-1} \geq p_{min}^{*}(1 - \frac{\alpha b_{o}}{5\alpha b_{o} + 4(1 + \frac{\alpha \ell}{5})}) \geq \frac{4}{5}p_{min}^{*}$$

And hence,

$$\min_{r} p_{\ell}(m) \geq 1 - (1 - \frac{4}{5}p_{min}^{*})^{m}$$

Now,

$$\beta^{t} \geq 2c'(\min_{r} p_{\ell}(m) - \sqrt{\alpha})$$

$$\geq 2c'(1 - (1 - \frac{4}{5}p_{min}^{*})^{m} - \sqrt{\alpha}) \geq \beta$$

where the last inequality is by our requirement on $c'$ and the fact that $1 - (1 - \frac{1}{5}p_{min}^{*})^{m} - \sqrt{\alpha} > 0$ by our requirement on $m$.

**Lemma 11.** Suppose $C^{*}$ is $(b_{o}, \alpha)$-stable. Then if we apply one step of Algorithm [11] with $m, c'$ satisfying conditions in Lemma [10], then conditioning on $\Omega_{t}$,

$$\Delta^{t} \leq \Delta^{*} - (1 - \frac{\beta}{t_{o} + i}) + \left[\frac{c'}{t_{o} + i}\right]^{2} \sum_{r} n_{r}^{\ell}||c_{r}^{\ell} - c_{r}^{*}||^{2}$$

$$+ 2c' \sum_{r} n_{r}^{\ell}(c_{r}^{\ell-1} - c_{r}^{*}, \xi_{r})$$

where $\xi_{r} := c_{r}^{*} - E[c_{r}^{*}|F_{t-1}]$.

**Proof.** Let $\Delta_{t}^{*} := n_{r}^{\ell}||c_{r}^{\ell} - c_{r}^{*}||^{2}$, so $\Delta^{t} = \sum_{r} \Delta_{r}^{t}$, and we use $p_{\ell}$ as a shorthand for $p_{\ell}(m)$. By the update rule of Algorithm [11]

$$\Delta_{t}^{r} = n_{r}^{\ell}||(1 - \eta^{r})(c_{r}^{\ell-1} - c_{r}^{*}) + \eta^{r}(c_{r}^{\ell} - c_{r}^{*})||^{2}$$

$$\leq n_{r}^{\ell}((1 - 2\eta^{r})||c_{r}^{\ell-1} - c_{r}^{*}||^{2} + 2\eta^{r}||c_{r}^{\ell-1} - c_{r}^{*}||^{2} + (\eta^{r})^{2}||c_{r}^{\ell-1} - c_{r}^{*}||^{2} + ||c_{r}^{\ell} - c_{r}^{*}||^{2})$$

Let $\xi_{r}^{t} = c_{r}^{*} - E[c_{r}^{*}|F_{t-1}]$, where

$$E[c_{r}^{*}|F_{t-1}] = (1 - p_{\ell})c_{r}^{\ell-1} + p_{\ell}m(A_{r}^{*})$$

Since

$$(c_{r}^{\ell-1} - c_{r}^{*}, \xi_{r}^{t}) = (c_{r}^{\ell-1} - c_{r}^{*}, E[c_{r}^{*}|F_{t-1}] + \xi_{r}^{t} - c_{r}^{*})$$

$$\leq (1 - p_{\ell})||c_{r}^{\ell-1} - c_{r}^{*}||^{2} + p_{\ell}||m(A_{r}^{*}) - c_{r}^{*}|| ||c_{r}^{\ell-1} - c_{r}^{*}||$$

We have

$$\Delta_{t}^{r} \leq n_{r}^{\ell} - 2\eta^{r}||c_{r}^{\ell-1} - c_{r}^{*}||^{2} - (1 - p_{\ell})||c_{r}^{\ell-1} - c_{r}^{*}||^{2}$$

$$+ (\eta^{r})^{2}||c_{r}^{\ell-1} - c_{r}^{*}||^{2} + ||c_{r}^{\ell} - c_{r}^{*}||^{2} + ||c_{r}^{\ell-1} - c_{r}^{*}||^{2}$$

Note

$$\sum_{r} n_{r}^{\star}||c_{r}^{\ell-1} - c_{r}^{*}||^{2} = \sum_{r} n_{r}^{\star}||m(A_{r}^{*}) - c_{r}^{*}||^{2}$$

$$\leq \sqrt{\sum_{r} n_{r}^{\star}||c_{r}^{\ell-1} - c_{r}^{*}||^{2}}(\sum_{r} n_{r}^{\star}||m(A_{r}^{*}) - c_{r}^{*}||^{2})$$

$$\leq \sqrt{\sum_{r} n_{r}^{\star}}||c_{r}^{\ell-1} - c_{r}^{*}||^{2} \leq \sqrt{\Delta^{*}} \leq \sqrt{\Delta^{*}}$$

where the first inequality is by Cauchy-Schwartz and the last inequality is by applying Lemma [11]. Finally, summing over $\Delta_{r}^{t}$, we get

$$\Delta^{t} = \sum_{r} \Delta_{r}^{t} \leq \Delta^{*} - (1 - \frac{\beta}{t_{o} + i}) + \left[\frac{c'}{t_{o} + i}\right]^{2} \sum_{r} n_{r}^{\ell}||c_{r}^{\ell} - c_{r}^{*}||^{2}$$

$$+ 2c' \sum_{r} n_{r}^{\ell}(c_{r}^{\ell-1} - c_{r}^{*}, \xi_{r})$$

The second inequality is by $\beta^{t} \geq \beta$, as proven in Lemma [10].
Lemma 12. Suppose $X$ satisfies (A1), $C^o \in \text{conv}(X)$, and $C^*$ is $(b_o, \alpha)$-stable. If we run one step of Algorithm 4 with $m, c'$ satisfying conditions in Lemma 7, then conditioning on $\Omega_i$, we have, for any $\lambda > 0$,

$$E_i\{\exp(\lambda \Delta^i)\} = \exp \left\{ \lambda ((1 - \beta \frac{t_o}{t_o + i})^\Delta_i - 1) + \frac{(c')^2 B}{(t_o + i)^2} + \frac{\lambda(c')^2 B^2}{2(t_o + i)^2} \right\}$$

Proof. By Lemma 24 we have (11) and (12) are both upper bounded by $B$. By Lemma 11 we have

$$E_i\{\exp(\lambda \Delta^i)\} = \exp \left\{ \lambda ((1 - \beta \frac{t_o}{t_o + i})^\Delta_i - 1) + \frac{(c')^2 B}{(t_o + i)^2} \right\}$$

Since

$$2\frac{\lambda c'}{i + t_o} \sum_r n_r^{\ast}(\xi_r, c_{r-1} - c_r') \leq 2\frac{\lambda c'}{i} \frac{\Delta_i}{t_o}$$

and $E_i\{\frac{2c'}{i + t_o} \sum_r n_r^{\ast}(\xi_r, c_{r-1} - c_r')|F_{i-1}\} = 0$, by Hoeffding's lemma

$$E_i\left\{ \exp \left\{ \frac{2c'}{i + t_o} \sum_r n_r^{\ast}(\xi_r, c_{r-1} - c_r')|F_{i-1}\right\} \right\} \leq \exp \left\{ \frac{\lambda^2 (c')^2 B^2}{2(i + t_o)^2} \right\}$$

Combining this with the previous bound completes the proof.

Lemma 13 (adapted from [30]). For any $\lambda > 0$, $E_i\{\exp(\lambda \Delta^{i-1})\} \leq E_i\{\exp(\lambda \Delta^{i-1})\}$. 

Proof. By our partitioning of the sample space, $\Omega_{i-1} = \Omega_i \cup (\Omega_{i-1} \setminus \Omega_i)$, and for any $\omega \in \Omega_i$ and $\omega' \in \Omega_{i-1} \setminus \Omega_i$, $\Delta_i^{i-1}(\omega) = b_o \phi^* < \Delta_i^{i-1}(\omega')$. Taking expectation over $\Omega_i$ and $\Omega_{i-1}$, we get $E_i\{\exp(\lambda \Delta^{i-1})\} \leq E_i\{\exp(\lambda \Delta^{i-1})\}$.

Proposition 2. Fix any $0 < \delta < \frac{1}{e}$. Suppose $C^*$ is $(b_o, \alpha)$-stable. If $\Delta^o \leq \frac{1}{2}b_o \phi^*$, and if

$$m > \frac{\ln(1 - \sqrt{\alpha})}{\ln(1 - \frac{\delta}{2}p_{\min}^*)}$$

$$c' > \frac{\beta}{2[1 - \sqrt{\alpha} - (1 - \frac{\delta}{2}p_{\min}^*)]|m|} \quad \text{with} \quad \beta \geq 2$$

$$t_o \geq 768(c')^2(1 + \frac{1}{b_o})^2t_0^2 \ln^2 \frac{1}{\delta}$$

Then

$$P(\Omega_\infty) \leq \delta$$

(here we used $\Delta^o$ instead of $\Delta^i$ and treat the starting time, the $i$-th iteration in Theorem 3 as the zeroth iteration for cleaner presentation).
Case 1: $B > \frac{\lambda B^2}{2}$. We get

$$\frac{1}{2} b_o \phi^* - (B + \frac{\lambda B^2}{2}) (t_o + i)^2 \geq \Delta$$

since $t_o \geq \frac{16(c')^2(b_o + 1)n}{b_o} = \frac{16(c')^2 B}{b_o \phi^*} \geq \frac{16(c')^2 B}{b_o \phi^*}$.

Case 2: $B \leq \frac{\lambda B^2}{2}$. We get

$$\frac{1}{2} b_o \phi^* - (B + \frac{\lambda B^2}{2}) (t_o + i)^2 \geq 2\Delta - \frac{1}{\Delta} \ln \left( \frac{(t_o + i)^2}{\delta} \right)$$

$$= 2\Delta - \frac{1}{\Delta} \ln \left( \frac{(t_o + i)^2}{\delta} \right)$$

$$\geq 2\Delta - \frac{1}{\Delta} \ln \left( \frac{(t_o + i)^2}{\delta} \right)$$

Now we show

$$\frac{1}{\Delta} \ln \left( \frac{(t_o + i)^2}{\delta} \right) \leq \Delta$$

Since

$$t_o + i \geq t_o \geq 768(c')^2 (1 + \frac{1}{b_o}) \ln 1 + \frac{1}{\delta}$$

$$= 48\frac{b_o B^2}{(2b_o \phi^*)^2} \ln 1 \frac{1}{\delta}$$

$$\ln \frac{1}{\delta} \geq 1, \text{ and } \frac{16(c')^2 B^2}{(2b_o \phi^*)^2} \geq \frac{1}{2}, \text{ we can apply Lemma 25 with }$$

$$\begin{align*}
\frac{1}{\Delta} \ln \left( \frac{(t_o + i)^2}{\delta} \right) & \geq \Delta \\
\text{That is, } \frac{1}{\Delta} \ln \left( \frac{(t_o + i)^2}{\delta} \right) & \leq \Delta. \\
\text{Thus, for both cases, } & 2\Delta - (B + \frac{\lambda B^2}{2}) (t_o + i)^2 \geq \Delta
\end{align*}$$

and hence,

$$Pr(\omega \in \Omega_i \setminus \Omega_{i+1}) \leq e^{-\frac{1}{\delta} (t_o (1+i)^2) \Delta} = \frac{\delta}{(t_o + i)^2}$$

Finally, we have

$$Pr(\cup_{i \geq 1} \Omega_i \setminus \Omega_{i+1}) \leq \sum_{i=1}^{\infty} Pr(\omega \in \Omega_i \setminus \Omega_{i+1}) \leq \delta$$

**Proof of Theorem 3**  Since the conditions in Proposition 2 holds for any $t > i$, we apply it and get

$$Pr(\Omega_t) \geq 1 - Pr(\cup_{i \geq 1} \Omega_i \setminus \Omega_{i+1}) \geq 1 - \delta$$

This proves the first statement. Taking expectation over $\Omega_t$ conditioning on filtration $F_{t-1}$ with respect to the inequality derived in Lemma 1, we get

$$E_t[\Delta^i | F_{t-1}] \leq \Delta^{t-1} (1 - \frac{\beta}{t_o + t}) + \frac{c'}{t_o + t} B$$

since $[12]$ is bounded by $B$ by Lemma 24 and since $E_t[\xi_r^i | F_{t-1}] = 0, \forall r \in [k]$. Taking total expectation over $\Omega_t$, we get

$$E_t[\Delta^i] \leq E_t[\Delta^{t-1} (1 - \frac{\beta}{t_o + t}) + \frac{c'}{t_o + t} B$$

$$\leq E_{t-1}[\Delta^{t-1}] (1 - \frac{\beta}{t_o + t}) + \frac{c'}{t_o + t} B$$

We can apply Lemma 26 by letting $u_r \leftarrow E_{t+i}[\Delta^{t+i}]$ (we temporarily change the notation [t] to $E_{t+i}[\Delta^{t+i}]$ to match the notation in Lemma 26), $o_r \leftarrow o_r + i$, $a \leftarrow \beta$, and $b \leftarrow (c')^2 B$

$$E_t[\Delta^i] \leq (t_o + i + 1) \Delta^i + \frac{(c')^2 B}{(t_o + t + 1)^{i+1}} \frac{1}{t_o + t + 1}$$

\square

10 Appendix D: Proofs of Theorem 1 and Theorem 2

One subtlety we need to point out before the proofs is that, in Algorithm 1, the learning rate $\eta^t$, as well as the update rule:

$$c_r^t \leftarrow (1 - \eta^t_r) c_r^{t-1} + \eta^t_r \hat{c}^t_r$$

is only defined for a cluster $r$ that is “sampled” at the $t$-th iteration. However, even if the cluster is not “sampled”, i.e., $c_r^t = c_r^{t-1}$, the same update rule with $c_r^t = c_r^{t-1}$ and the same learning rate still holds for this case. So in our analysis, we equivalently treat each cluster $r$ as updated with learning rate $\eta_r^t$ and differentiate between a sampled and not-sampled cluster only through the definition of $\hat{c_r^t}$.

**Proof leading to Theorem 1**

**Lemma 14.** Suppose $\forall r \in [k], \eta_r^t \leq \eta_{\max}$ w.p. 1. Then, $E_t[\phi^{t+1} - \phi^t | F_t] \leq -2 \min_{r \in [k]} \eta_r^{t+1} \phi_r^{t+1} + (\eta_{\max})^2 6 \phi^t$, where $\phi^t := \sum_r \sum_{x \in A_r^{t+1}} ||x - \mu(A_r^{t+1})||^2$.

**Proof of Lemma 14.** For simplicity, we denote $E_t[|F_t|]$ by $E_t[|\cdot|]$ (the same notation is also used as a shorthand to $E_t[\Omega_i]$ in the proof of Theorem 3; we abuse the notation here).

$$E_t[\phi^{t+1}] = E_t[\sum_{r=1}^{k} \sum_{x \in A_r^{t+1}} ||x - c_r^{t+1}||^2]$$

$$\leq E_t[\sum_{r \in A_r^{t+1}} ||x - c_r^{t+1}||^2]$$

$$= E_t[\sum_{r \in A_r^{t+1}} \sum_{x \in A_r^{t+1}} (1 - \eta_r^{t+1})^2 ||x - c_r^{t+1}||^2$$

$$+(\eta_r^{t+1})^2 ||x - c_r^{t+1}||^2 + 2 \eta_r^{t+1} (1 - \eta_r^{t+1}) (x - c_r^{t+1})]$$

where the inequality is due to the optimality of clustering $A_r^{t+1}$ for centroids $C^{t+1}$. Since

$$E_t[\hat{c}_{r}^{t+1}] = (1 - \eta_r^{t+1}) c_r^{t+1} + \eta_r^{t+1} \mu(A_r^{t+1})$$
we have

\[ \langle x - c_t, x - c_t^{+1} \rangle = (1 - p_t^{+1})\|x - c_t\|^2 + p_t^{+1} \langle x - c_t, x - m(A_t^{+1}) \rangle \]

Plug this into the previous inequality, we get

\[ E_t[\phi^{+1}] \leq \sum_r (1 - 2\eta_r^{+1})\phi_t^r + (\eta_r^{+1})^2 \phi_t^r + p_t^{+1} \{ (1 - p_r^{+1}) \sum_{x \in A_t^{+1}} \|x - c_r\|^2 \]

\[ + 2\eta_r^{+1} \{ (1 - p_r^{+1}) \sum_{x \in A_t^{+1}} \|x - c_r\|^2 \}

\[ + p_r^{+1} \sum_{x \in A_t^{+1}} \langle x - c_r, x - m(A_t^{+1}) \rangle \}

\[ = \phi_t^r - 2 \sum_r \eta_r^{+1} p_r^{+1} \phi_t^r
\]

\[ + 2 \sum_r \eta_r^{+1} p_r^{+1} \sum_{x \in A_t^{+1}} \langle x - c_r, x - m(A_t^{+1}) \rangle \}

\[ + (\eta_r^{+1})^2 \phi_t^r + (\eta_r^{+1})^2 \sum_{x \in A_t^{+1}} \|x - c_r^{+1}\|^2 \]

Now,

\[ \sum_{x \in A_t^{+1}} \langle x - c_r, x - m(A_t^{+1}) \rangle \]

\[ = \sum_{x \in A_t^{+1}} \langle x - m(A_t^{+1}) + m(A_t^{+1}) - c_r, x - m(A_t^{+1}) \rangle \]

\[ = \sum_{x \in A_t^{+1}} \|x - m(A_t^{+1})\|^2 + \sum_{x \in A_t^{+1}} \langle m(A_t^{+1}) - c_r, x - m(A_t^{+1}) \rangle \]

since \( x \in A_t^{+1} \) for all \( r \), by property of the mean of a cluster. Then

\[ E_t[\phi^{+1}] \leq \phi_t^r + 2 \sum_r \eta_r^{+1} p_r^{+1} (-\phi_t^r + \hat{\phi}_t^r)
\]

\[ + (\eta_r^{+1})^2 [\phi_t^r + E_t[\sum_{x \in A_t^{+1}} \|x - c_r^{+1}\|^2] \]

Now a key observation is that \( p_r^{+1} = 0 \) if and only if cluster \( A_r^{+1} \) is empty, i.e., degenerate. Since the degenerate clusters do not contribute to the k-means cost, we have \( \sum_{r \in [k], p_r^{+1} > 0} \phi_t^r = \phi_t^r \), and similarly, \( \sum_{r \in [k], p_r^{+1} > 0} \hat{\phi}_t^r = \hat{\phi}_t^r \). Therefore,

\[ E_t[\phi^{+1}] \leq \phi_t^r - 2 \sum_{r \in [k], p_r^{+1} > 0} \eta_r^{+1} p_r^{+1} (-\phi_t^r + \hat{\phi}_t^r)
\]

\[ + (\eta_{\max}^{+1})^2 (E_t[\sum_{x \in A_t^{+1}} \|x - c_r^{+1}\|^2 + \phi_t^r] \]

\[ = \phi_t^r - 2 \sum_{r \in [k], p_r^{+1} > 0} \eta_r^{+1} p_r^{+1} (-\phi_t^r + \hat{\phi}_t^r) + (\eta_{\max}^{+1})^2 2\phi_t^r \]

where the last inequality is by Lemma 23.

\[ \text{Lemma 15. Suppose Assumption (A) holds. If we run Algorithm 1 on X with } \eta_t^* = \frac{1}{\eta_t}, \text{ and } t_o > 1, \text{ with any initial set of } k \text{ centroids } C^0 \in \text{conv}(X). \text{ Then for any } \delta > 0, \exists s.t. \Delta(C^t, C^*) \leq \delta \text{ with } C^* := m(A^*) \text{ for some } A^* \in \{A^*_k\}. \]

**Proof of Lemma 15.** First note that since \( \{C^*_k\} \) includes all stationary points with \( 1 \leq k' \leq k \) non-degenerate centroids, and at any time \( t \), \( C^t \) must have \( k' \leq k \) non-degenerate centroids, so there exists \( C^t \in \{C^*_k\}_{k} \) such that \( \Delta(C^t, C^*) \) is well defined. For a contradiction, suppose \( \forall t \geq 1, \Delta(C^t, C^*) > \delta \), for all \( C^t \in \{C^*_k\}_{k} \). Then

**Case 1:** \( m(A_t^{+1}) \in \{C^*_k\} \)

Then

\[ \Delta(C^t, m(A_t^{+1})) > \delta \]

by our assumption.

**Case 2:** \( m(A_t^{+1}) \notin \{C^*_k\} \)

Since \( C^t \in \text{Cl}(v^{-1}(A_t^{+1})) \) by our definition, applying Lemma 2,

\[ \Delta(C^t, m(A_t^{+1})) > \min\{\delta, r_{min}(m(A_t^{+1}))\} \]

So for both cases,

\[ \Delta(C^t, m(A_t^{+1})) > \min\{\delta, r_{min}(m(A_t^{+1})) \}
\]

Let denote \( \delta_o := \min\{\delta, r_{min}(m(A_t^{+1})) \} \), then by Lemma 14

\[ \min_{x \in [k], p_x^{+1} > 0} p_x^{+1} \geq 1 - \frac{\delta_o}{c^t} \]

Also note

\[ \phi_t^r - \hat{\phi_t^r} = \sum_{r \in [k], x \in A_t^{+1}} \|x - c_r\|^2 - \|x - m(A_t^{+1})\|^2 \]

\[ = \sum_{r \in [k], x \in A_t^{+1}} \|c_r - m(A_t^{+1})\|^2 + \eta_{\max}^{+1} = \Delta(C^t, m(A_t^{+1})) \geq \delta_o \]

Then \( \forall t \geq 1, \)

\[ F_t[\phi^{+1}] - F_t[\phi^0] \]

\[ \leq -2 \eta_{\max}^{+1} \|c_r - m(A_t^{+1})\|^2 \frac{\delta_o}{t + 1 + t_o} + \frac{6\phi_{\max}(c)^2}{(t + 1 + t_o)^2} \]

Summing up all inequalities,

\[ F_t[\phi^{+1}] - F_t[\phi^0] \]

\[ \leq -2 \eta_{\max}^{+1} \|c_r - m(A_t^{+1})\|^2 \frac{\delta_o}{t + 1 + t_o} + \frac{6\phi_{\max}(c)^2}{(t + 1 + t_o)^2} \]
Since \( t \) is unbounded and \( \ln \frac{\ln t + 1}{2t\Delta} \) increases with \( t \) while \( \frac{\ln(\Delta t)}{t} \) is a constant, \( \exists T \) such that for all \( t \geq T \), \( E\phi_t - \phi_0 \leq -\phi_0 \), which means \( E\phi_t \leq 0 \), for all \( t \) large enough. This implies the \( k \)-means cost of some clusterings is negative, which is impossible. So we have a contradiction. 

**Proof setup of Theorem 1.** The goal of the proof is to show that first, with high probability, the algorithm converges to some stationary clustering, \( A^* \in \{A^*_i\}_k \). We call this event \( G \); formally,

\[
G := \{ \exists T \geq 1, \exists A^* \in \{A^*_i\}_k \text{ s.t. } A^i = A^* , \forall t \geq T \}
\]

Second, we want to establish the \( O(1) \) expected convergence rate of the algorithm to this stationary clustering \( A^* \).

To prove that the event \( G \) has high probability, we first consider random variable \( \tau \):

\[
\tau := \min\{ t \geq 0 \mid \min_{A^i \in \{A^*_i\}_k} \Delta(C^i, m(A^*)) \leq \frac{1}{2} r_{\min} \phi^* \}
\]

That is, \( \tau \) is the first time the algorithm “hits” a stationary clustering; \( \tau \) is a stopping time since \( \forall t \geq 0, \{ \tau \leq t \} \) is \( F_t \)-measurable. By Lemma 15,

\[
Pr(\{ \tau < \infty \}) = Pr(\{ \tau \in \mathbb{N} \}) = Pr(\cup_{t \geq 0} (\{ \tau = T \})) = 1
\]

Fixing \( \tau \), we denote the stationary clustering that the algorithm “hits” by

\[
A^*(\tau) := \arg\min_{A^i \in \{A^*_i\}_k} \Delta(C^i, m(A^*))
\]

\( A^*(\tau) \) is well defined; the reason is when \( \Delta(C^i, m(A^*)) \leq \frac{1}{2} r_{\min} \phi^* \), \( A^i = A^* \), so there can be only one minimizer.

We will prove a subset \( G_{\phi} \subseteq G \) holds with high probability.

To do this, we construct \( G_{\phi} \) as a union of disjoint events determined by the realization of \( \tau \) and \( A^*(\tau) \): we define events

\[
G_T(A^*) := \{ \tau = T \} \cap \{ A^*(\tau) = A^* \} \cap \{ \forall t \geq 0, \Delta^i \leq r_{\min} \phi^* \}
\]

Then we can represent the event where the algorithm’s iterate converges to a particular stationary clustering \( A^* \) as

\[
G(A^*) := \cup_{T \geq 0} G_T(A^*)
\]

Finally, we define

\[
G_o := \cup_{A^* \in \{A^*_i\}_k} G(A^*)
\]

\( G_o \subseteq G \) since the event \( \Delta^i \leq r_{\min} \phi^* \) implies \( A^i = A^* \).

**Proof of Theorem 7.** Fix any \( (T, A^*) \), conditioning on \( \{ \tau = T \} \cap \{ A^*(\tau) = A^* \} \), since we have

\[
c' > \frac{\phi_{\text{max}}}{(1 - e^{-\frac{\phi}{\min} \phi_{\text{opt}}})}
\]

We can invoke Lemma 16 to get \( \forall t < T \),

\[
E[\phi_t - \phi(A^*)]G_T(A^*) = O\left( \frac{1}{t} \right)
\]

Now let’s consider the case \( t \geq T \). Since by Lemma 2, \( A^* \) is \( (r_{\min}, 0) \)-stable, we can apply Theorem 3 in this context, the parameters in the statement of Theorem 3 are \( \phi_o = r_{\min}, \alpha = 0, p_{\min} \geq \frac{\phi_{\text{opt}} - \phi}{\min} \).

Therefore, for any

\[
m \geq 1
\]

\[
c' > \frac{\beta}{2(1 - e^{-\frac{\phi}{\min} \phi_{\text{opt}}})} \quad \text{with } \beta \geq 2
\]

and

\[
t_o \geq 768(c')^2(1 + \frac{1}{r_{\min}})^2 n^2 \ln \frac{1}{\delta}
\]

the conditions required by Theorem 3 are satisfied. Then by the first statement of Theorem 3

\[
Pr(\{ \forall t \geq 0, \Delta^i \leq r_{\min} \phi^* \}) \leq E\Delta(C^i, m(A^*)) \leq E\Delta(C^i, \phi^*) = O\left( \frac{1}{t} \right)
\]

Finally, we turn to prove \( Pr(G) \) is large. Recall

\[
Pr(G) \geq Pr(G_o) = Pr(\cup_{T \geq 0} \cup_{A^* \in \{A^*_i\}_k} G_T(A^*) \cup \cup_{T \geq 0} G_T(A^*))
\]

where the second equality holds because the events \( G_T(A^*) \) are disjoint for different pairs of \( (T, A^*) \), since the stopping time \( \tau \) and the minimizer \( A^*(\tau) \) are unique for each experiment. Since

\[
\sum_{T \geq 0} \sum_{A^* \in \{A^*_i\}_k} Pr(G_T(A^*))
\]

\[
= \sum_{T, A^*} Pr(\{ \tau = T \} \cap \{ A^*(\tau) = A^* \})
\]

\[
\geq (1 - \delta) \sum_{T, A^*} Pr(\{ \tau = T \} \cap \{ A^*(\tau) = A^* \})
\]

\[
= (1 - \delta) Pr(\{ \cup_{T \geq 0} \cup_{A^* \in \{A^*_i\}_k} \tau = T \} \cap \{ A^*(\tau) = A^* \})
\]

\[
= (1 - \delta) Pr(\{ \cup_{T \geq 0} \{ \tau = T \} \cap \{ A^*(\tau) = A^* \})
\]

where the inequality is by (15), and the last two equalities are due to the finiteness of \( \{A^*_i\}_k \) and by (13), respectively.
which proves the second statement. Finally, combining
inequalities \[14\] and \[16\], we have \(\forall t \geq 1\) and \(\forall t \geq 1\),
\[E\{\phi^i - \phi(A^+)|G_T(A^+)\} = O\left(\frac{1}{t}\right)\]
Since the quantity \(\phi^i - \phi(A^+)\) is independent of \(T\), we
reach the conclusion
\[E\{\phi^i - \phi(A^+)|G(A^+)\} = O\left(\frac{1}{t}\right)\]

**Lemma 16.** Suppose the assumptions and settings in
Theorem \[1\] hold, conditioning on any \(G_T(A^+)\), we have \(\forall t \geq 1\),
\[E\{\phi^i - \phi(A^+)|G_T(A^+)\} = O\left(\frac{1}{t}\right)\]

**Proof.** First observe that conditioning on the event \(G_T(A^+), \Delta(C^i, C^o) > \frac{1}{2}\min \phi^e, \forall t < T\). Now we are in a setup
similar to that in the proof Lemma \[15\] and the argument therein will lead us to the conclusion that
\[\phi^i - \hat{\phi}^i > \min\left\{\frac{1}{2}\min \phi^e, \min \phi^e\right\} = \frac{1}{2}\min \phi^e\]
Proceeding as in Lemma \[15\] we have conditioning on \(G_T(A^+), \Delta(C^i, C^o) > \frac{1}{2}\min \phi^e, \forall t < T\), we use \(p_{\min} := \min_{r \in [k]} \frac{n}{2}\) to characterize the fraction of the smallest cluster in \(A^o\) to the
entire dataset. We use \(w_r := \max_{s \in A^o} |x - c_r|\) to characterize the ratio between average and maximal “spread” of
cluster \(A^o_r\), and we let \(w_{\min} := \min_{r \in [k]} w_r\).

**10.1 Existence of stable stationary point**
under geometric assumptions on the dataset

First, we observe that our Assumption (B) implies two lower bounds on \(\|c^o_r - c^i_r\|, \forall r, s \neq r\). Let \(x \in A^o_r \cap A^o_s\). Split \(x\) into its projection on the line joining \(c^i_r\) and \(c^o_r\), and its orthogonal component:
\[x = \frac{1}{2}(c^o_r + c^i_r) + \lambda(c^o_r - c^i_r) + u\]
with \(u \perp c^o_r - c^i_r\). Note \(\lambda\) measures the ratio between departure of the projected point from the mid-point of \(c^o_r\) and \(c^i_r\) and the norm \(\|c^o_r - c^i_r\|\). By minimality of our
definition of margin \(\Delta_{rs}\),
\[\|x - \frac{1}{2}(c^o_r + c^i_r)\| = \lambda\|c^o_r - c^i_r\| \geq \frac{1}{2}\Delta_{rs}\]
In addition, since \(c^i_r\) is the mean of \(A^o_r\), we know there exists \(x \in A^o_r\) such that \(\tilde{x}\) falls outside of the line segment \(c^o_r - c^i_r\) (or exactly on \(c^i_r\) in the special case where all points
projects on \(c^i_r\)). Similar holds for \(c^o_r\). Thus,
\[\|c^o_r - c^i_r\| \geq \Delta_{rs} \geq f(\alpha)\sqrt{\phi^e(\frac{1}{\sqrt{n_r}} + \frac{1}{\sqrt{n_s}})}\]

**Lemma 17 (Theorem 5.4 of [12].)** Suppose \((X, C^*)\)
satisfies (B). If \(\forall r, \Delta^r + \Delta^s \leq \frac{\Delta_{rs}}{16}\). Then for any \(s \neq r, |A^o_r \cap A^o_s| \leq \frac{b^2}{f(\alpha)}, \) where \(b := \max_{r,s} \frac{\Delta^r + \Delta^s}{\Delta_{rs}}\).

The proof is almost verbatim of Theorem 5.4 of [12]; we include it here for completeness.

**Proof.** Since the projection of \(x\) on the line joining \(c^i_r, c^o_r\) is closer to \(s\), we have
\[x(c^i_r - c^o_r) \geq \frac{1}{2}(c^i_r - c^o_r)(c^o_r + c^i_r)\]
Substituting (17) into the inequality above,
\[
\frac{1}{2}(c_r^* + c_r')(c_r' - c_r) + \lambda(c_r^* - c_r')(c_r' - c_r')
+ u(c_r' - c_r') \geq \frac{1}{2}(c_r' - c_r')(c_r' + c_r')
\] (20)
Since \( u \perp c_r^* - c_r^* \), let \( \Delta = \Delta_r^* + \Delta_r' \). We have
\[
u(c_r' - c_r') = u(c_r' - c_r^* - (c_r' - c_r')) \leq \|u\| \Delta
\]
Rearranging (20), we have
\[
\frac{1}{2}(c_r^* + c_r' - c_r^* - c_r')(c_r' - c_r')
+ \lambda(c_r^* - c_r')(c_r' - c_r') + u(c_r' - c_r') \geq 0
\]
\[
\geq \frac{\Delta^2}{2} + \frac{\Delta}{2} \|c_r^* - c_r^*\|^2 - \lambda \|c_r^* - c_r^*\|^2
+ \lambda \Delta \|c_r^* - c_r^*\| + \|u\| \Delta \geq 0
\]
Therefore,
\[
\|x - c_r^*\| = \| \frac{1}{2} - \lambda \|c_r^* - c_r^*\|^2 - \frac{\Delta^2}{2} + \frac{\Delta}{2} \|c_r^* - c_r^*\|^2 - \lambda \|c_r^* - c_r^*\|^2 \geq \frac{\Delta^2}{64\Delta}
\]
where the last inequality is by our assumption that \( \Delta \leq \frac{\Delta^2}{64\Delta} \),
and \( \lambda \geq \frac{\Delta^2}{64\Delta} \) by (18). By previous inequality and our assumption
on \( f \), for all \( s \not= r \)
\[
|A_s^* \cap A_r^*| \leq \frac{\Delta^2}{64\Delta} \|c_r^* - c_s^*\| \leq \sum_{x \in A_s^* \cap A_r^*} \|x - c_r^*\|^2
\]
So \( |A_s^* \cap A_r^*| \leq \frac{\Delta^2}{64\Delta} \|x - c_r^*\|^2 \leq \frac{\Delta^2}{64\Delta} \|x - c_r^*\|^2 \), where the second inequality is by (19).
That is, \( \frac{|A_s^* \cap A_r^*|}{\|x - c_r^*\|^2} \leq \frac{\Delta^2}{64\Delta} \). Similarly, for all \( s \not= r \),
\[
\frac{|A_s^* \cap A_r^*|}{\|x - c_r^*\|^2} \leq \frac{\Delta^2}{64\Delta} \|x - c_r^*\|^2
\]
Summing over all \( s \not= r \),
\[
\frac{|A_s^* \cap A_r^*|}{\|x - c_r^*\|^2} = \rho_{out} + \rho_{in} \leq \frac{\Delta^2}{64\Delta} \frac{\|x - c_r^*\|^2}{\|x - c_r^*\|^2} = \frac{k^2}{T}
\]
Lemma 18. Fix a stationary point \( C^* \) with \( k \) centroids, and any other set of \( k' \) centroids, \( C' \), with \( k' \geq k \) so that \( C \) has exactly \( k \) non-degenerate centroids. We have
\[
\phi(C) - \phi^* \leq \min_s \sum_r n_r^* \|c_s - c_r^*\|^2 = \Delta(C, C^*)
\]
Proof. Since degenerate centroids do not contribute to \( k \)-means cost, in the following we only consider the sets of non-degenerate centroids \( \{c_s, s \in [k]\} \subset C \) and \( \{c_r^*, r \in [k]\} \subset C^* \). We have for any permutation \( \pi \),
\[
\phi(C) - \phi^* = \sum_{s \in A_s} \|x - c_s\|^2 - \sum_{r \in A_r^*} \|x - c_r^*\|^2
\]
\[
\leq \sum_r \sum_{s \in A_s} \|x - c_{\pi(s)}\|^2 - \sum_r \sum_{r \in A_r^*} \|x - c_r^*\|^2
\]
\[
= \sum_r n_r^* \|c_{\pi(s)} - c_r^*\|^2
\]
where the last inequality is by optimality of clustering based on Voronoi diagram, and the second inequality is by applying the centroidal property in Lemma 21 to each centroid in \( C^* \). Since the inequality holds for any \( \pi \), it must holds for \( \min \sum_r n_r^* \|c_{\pi(s)} - c_r^*\|^2 \), which completes the proof. \( \blacksquare \)

Proofs regarding seeding guarantee

Lemma 19 (Theorem 4 of [19]). Suppose \((X, C^*)\) satisfies (B). If we obtain seeds from Algorithm 3 then
\[
\Delta(C^0, C^*) \leq \frac{1}{2} f(\alpha)^2 \phi^*
\]
with probability at least \( 1 - m_o \exp(-2(f(\alpha) - 1)^2 w_{min}^2) - k \exp(-m_o \phi_{min}) \).

Proof. First recall that, as in [19], assumption (B) implies center-separability assumption in Definition 1 of [19], i.e.
\[
\forall r \in [k], s \not= r, \|c_r^* - c_s^*\| \geq f(\alpha) \sqrt{\phi^*} \leq \sqrt{f(\alpha)} \sqrt{\phi^*} \geq \sqrt{\phi^*}
\]
where \( f(\alpha) \geq \max_{r \in [k], s \not= r} \frac{n_r^*}{n_s^*} \). Applying Theorem 4 of [19] with \( \mu_r = c_r^* \) and \( \nu_r = c_r^* \), we get \( \forall r \in [k], \|c_r^* - c_r^*\| \leq \sqrt{f(\alpha)} \sqrt{\phi^*} \) with probability at least \( 1 - m_o \exp(-2(f(\alpha) - 1)^2 w_{min}^2) - k \exp(-m_o \phi_{min}) \). Summing over all \( r \), the previous event implies \( \sum_r n_r^* \|c_r^* - c_r^*\|^2 \leq f(\alpha)^2 \phi^* \). Therefore the last inequality is by the assumption that \( f \geq 64^2 \) in (B). \( \blacksquare \)

Lemma 20. Assume the conditions Lemma 19 hold. For \( \alpha > 0 \), if in addition,
\[
f(\alpha) \geq 5 \left( \frac{1}{2w_{min}} \ln \frac{2}{\epsilon_0} \ln \frac{2k}{\epsilon} \right)
\]
If we obtain seeds from Algorithm 3 choosing
\[
\frac{\ln 2k}{\epsilon} \leq m_o \leq \frac{\epsilon}{2} \exp(2(f(\alpha)/4 - 1)^2 w_{min}^2)
\]
Then \( \Delta(C^0, C^*) \leq \frac{1}{2} f(\alpha)^2 \phi^* \) with probability at least \( 1 - \xi \).

Proof. By Lemma 19, a sufficient condition for the success probability to be at least \( 1 - \xi \) is:
\[
m_o \exp(-2(f(\alpha) - 1)^2 w_{min}^2) \leq \frac{\xi}{2}
\]
and
\[
k \exp(-m_o \phi_{min}) \leq \frac{\xi}{2}
\]
This translates to requiring
\[
\frac{1}{\phi_{min}} \ln \frac{2k}{\epsilon} \leq m_o \leq \frac{\epsilon}{2} \exp(2(f(\alpha)/4 - 1)^2 w_{min}^2)
\]
\( \epsilon \) note: “\( \alpha \)” in [19] is defined as \( \min_{s \in [k], s \not= r} \frac{n_r^*}{n_s^*} \), which is not to be confused with our “\( \alpha \)”. 

Cheng Tang, Claire Monteleoni
Note for this inequality to be possible, we also need \( \frac{1}{\Delta \min} \ln \frac{2}{\xi} \leq \frac{\xi}{2} \exp(2(\frac{f(\alpha)}{4}) - 1)^2 \omega_{\min}^2) \), imposing a constraint on \( f(\alpha) \). Taking logarithm on both sides and rearrange, we get

\[
\left( \frac{f(\alpha)}{4} - 1 \right)^2 \geq \frac{1}{2 \omega_{\min}} \ln(2 \phi_{\min} \ln \frac{2k}{\xi})
\]

This satisfied since \( f(\alpha) \geq 5 \sqrt{\frac{1}{2 \omega_{\min}} \ln(2 \phi_{\min} \ln \frac{2k}{\xi})} \).

**Proof of Theorem 2** By Proposition 1 \((X, C^*)\) satisfying (B) implies \( C^* = (\frac{f(\alpha)}{4}, \alpha)\)-stable. Let \( b_0 := (\frac{f(\alpha)}{4})^2 \), and we denote event \( F := (\Delta(C^0, C^*) \leq \frac{\xi}{2} b_0 \phi^2) \). Since

\[
f(\alpha) \geq 5 \sqrt{\frac{1}{2 \omega_{\min}} \ln(2 \phi_{\min} \ln \frac{2k}{\xi})}, \quad \text{and} \quad \phi_{\min} < m_a < \frac{\xi}{2} \exp(2(\frac{f(\alpha)}{4}) - 1)^2 \omega_{\min}^2),
\]

we can apply Lemma 20 to get

\[
Pr\{F\} \geq 1 - \xi
\]

Conditioning on \( F \), we can invoke Theorem 3, since (A1) is satisfied implicitly by (B), \( C^* \subset \text{conv}(X) \) by the sampling method used in Algorithm 2, and we can guarantee that the setting of our parameters, \( m, c', \) and \( t_0 \), satisfies the condition required in Theorem 3. Let \( \Omega_1 \) be as defined in the main paper, by Theorem 3 \( \forall t \geq 1, \)

\[
E\{\Delta_1^t|\Omega_1, F\} = O\left(\frac{1}{t}\right) \quad \text{and} \quad Pr\{\Omega_1|F\} \geq 1 - \delta
\]

So

\[
Pr\{\Omega_1 \cap F\} = Pr\{\Omega_1|F\} Pr\{F\} \geq (1 - \delta)(1 - \xi)
\]

Finally, using Lemma 18 and letting \( G_t := \Omega_1 \cap F \), we get the desired result.

11 Appendix E: auxiliary lemmas

**Equivalence of Algorithm 1 to stochastic k-means** Here, we formally show that Algorithm 1 with specific instantiation of sample size \( m \) and learning rates \( \eta_i \) is equivalent to online k-means \([8]\) and mini-batch k-means \([13]\).

**Claim 1.** In Algorithm 1, if we set a counter for \( N_i := \sum_{t=1}^{\tau_i} n_i \) and if we set the learning rate \( \eta_i := \frac{n_i}{N_i^t} \), then provided the same random sampling scheme is used.

1. When mini-batch size \( m = 1 \), the update of Algorithm 7 is equivalent to that described in Section 3.3, \([8]\).
2. When \( m > 1 \), the update of Algorithm 7 is equivalent to that described from line 3 to line 14 in Algorithm 1, \([13]\) with mini-batch size \( m \).

**Proof.** For the first claim, we first re-define the variables used in Section 3.3, \([8]\). We substitute index \( k \) in \([8]\) with \( r \) used in Algorithm 1. For any iteration \( t \), we define the equivalence of definitions: \( s \leftarrow x_i, c'_i \leftarrow w_k, n_i \leftarrow \Delta n_k, N_i^t \leftarrow n_k \). According to the update rule in \([8]\), \( \Delta n_k = 1 \) if the sampled point \( x_i \) is assigned to cluster with center \( w_k \). Therefore, the update of the \( k \)-th centroid according to online k-means in \([8]\) is:

\[
w_k \leftarrow w_k + \frac{1}{n_k}(x_i - w_k)1_{\{\Delta n_k = 1\}}
\]

Using the re-defined variables, at iteration \( t \), this is equivalent to

\[
c_{t}^i = c_{t-1}^i + \frac{1}{N_i^t}(s - c_{t-1}^i)1_{\{s \neq 0\}}
\]

Now the update defined by Algorithm 1 with \( m = 1 \) and \( \eta_i = \frac{n_i}{N_i^t} \) is:

\[
c_t^i = c_t^{i-1} + \frac{n_i}{N_i^t}(s - c_t^{i-1})1_{\{s \neq 0\}} = c_t^{i-1} + \frac{n_i}{N_i^t}(s - c_t^{i-1})1_{\{s \neq 0\}}
\]

since \( n_i^t \) can only take value from \( \{0, 1\} \). This completes the first claim.

For the second claim, consider line 4 to line 14 in Algorithm 1, \([13]\). We substitute their index of time \( i \) with \( t \) in Algorithm 1. We define the equivalence of definitions: \( m \leftarrow b, S^t \leftarrow \mathcal{M}_t, s \leftarrow x, c_{t+1}^i \leftarrow d(x), c_{t+1}^i \leftarrow c \).

At iteration \( t \), we let \( v[s_{t-1}^i] \) denote the value of counter \( v \) upon completion of the loop from line 9 to line 14 for each center \( c_i \), then \( N_i^t \leftarrow v[s_{t-1}^i] \). Since according to Lemma 22 from line 9 to line 14, the updated centroid \( c_i \) after iteration \( t \) is

\[
c_t^i = \frac{1}{v[s_{t-1}^i]} \sum_{s \in c_{t-1}^i \cap S_{t-1}^i} s = \frac{1}{N_i^t} \sum_{s \in c_{t-1}^i \cap S_{t-1}^i} s
\]

This implies

\[
c_t^i - c_{t-1}^i = \frac{1}{N_i^t} \sum_{s \in c_{t-1}^i \cap S_{t-1}^i} s - c_{t-1}^i
\]

\[
= \frac{1}{N_i^t} \sum_{s \in S_{t-1}^i} s + \sum_{s' \in c_{t-1}^i \cap S_{t-1}^i} s' - c_{t-1}^i
\]

\[
= \frac{1}{N_i^t} \sum_{s \in S_{t-1}^i} s + N_i^t \cdot c_{t-1}^i - c_{t-1}^i
\]

\[
= \frac{1}{N_i^t} \sum_{s \in S_{t-1}^i} s - c_{t-1}^i + \frac{n_i}{N_i^t} c_{t-1}^i
\]

Hence, the updates in Algorithm 1 and line 4 to line 14 in Algorithm 1, \([13]\) are equivalent.

**Lemma 21.** (Centroidal property, Lemma 2.1 of \([11]\)). For any point set \( Y \) and any point \( c \) in \( \mathbb{R}^d \),

\[
\sum_{s \in Y} \|x - c\|^2 = \sum_{s \in Y} \|x - m(Y)\|^2 + \|Y\||m(Y) - c\|^2
\]

**Lemma 22.** Let \( w_t, g_t \) denote vectors of dimension \( \mathbb{R}^d \) at time \( t \). If we choose \( w_0 \) arbitrarily, and for \( t = 1 \ldots T \), we repeatedly apply the following update

\[
w_t = (1 - \frac{1}{t})w_{t-1} + \frac{1}{t} g_t
\]

Then

\[
w_T = \frac{1}{T} \sum_{t=1}^{T} g_t
\]
Proof. We prove by induction on \( T \). For \( T = 1 \), \( w_1 = (1 - 1)w_0 + g_1 = \frac{1}{2} \sum_{t=1}^T g_t \). So the claim holds for \( T = 1 \).

Suppose the claim holds for \( T \), then for \( T+1 \), by the update rule
\[
w_{T+1} = \left(1 - \frac{1}{T+1}\right)w_T + \frac{1}{T+1} \sum_{t=1}^T g_t
\]
\[
= \left(1 - \frac{1}{T+1}\right) T \sum_{t=1}^T g_t + \frac{1}{T+1} \sum_{t=1}^T g_t
\]
\[
= T \sum_{t=1}^T g_t + \frac{1}{T+1} \sum_{t=1}^T g_t
\]
\[
= \frac{T}{T+1} \sum_{t=1}^T g_t
\]
\[
= \frac{1}{T+1} \sum_{t=1}^{T+1} g_t
\]

So the claim holds for any \( T \geq 1 \). \( \square \)

**Lemma 23.** \( \forall t \geq 1 \), conditioning on \( F_t \), the noise term \( \phi \) is upper bounded by \( B_1 := 5 \hat{\phi}^* \).

**Proof.** Since
\[
\|x - c^{t+1}_r\|^2 \leq 2\|x - c^*_r\|^2 + 2\|c^*_r - c^{t+1}_r\|^2
\]
We have
\[
E[\sum_{r \in A_{t+1}^*} \|x - c^{t+1}_r\|^2 + \phi^* | F_t]
\]
\[
\leq 2 \sum_{r \in A_{t+1}^*} \|x - c^*_r\|^2 + \|c^*_r - c^{t+1}_r\|^2 + \phi^* \phi^*_t
\]

Now,
\[
E[\|c^*_r - c^{t+1}_r\|^2 | F_t] \leq \frac{\sum_{s \in S_t} \|c^*_r - s\|^2}{|S_t|} = \frac{\phi^*_r}{n^*_r}
\]
where \( S_t \) is the sampled from \( A_t \) in Algorithm \( \#1 \) and the inequality is by convexity of \( l_2 \)-norm. Substituting this into the previous inequality completes the proof. \( \square \)

**Lemma 24.** Suppose \( C^* \) is \( (b_0, \alpha) \)-stable. Conditioning on \( \Omega_i \), we have, The terms \( [11] \), and \( [12] \), for \( t = i \), are upper bounded by \( B := 4(b_0 + 1)n\hat{\phi}^* \).

**Proof.** Conditioning on \( \Omega_i \),
\[
\Delta^{i-1} \leq b_0 \phi^*
\]
By Lemma \( \#18 \) we also have
\[
\phi^i - \phi^* \leq \Delta^{i-1} \leq b_0 \phi^*
\]
By Cauchy-Schwarz,
\[
\sum_{r} n^*_r (c^*_r - c^i_r, c^i_r - c^i_r) \leq \sqrt{\sum_{r} n^*_r \|c^*_r - c^i_r\|^2 \sum_{r} n^*_r \|c^i_r - c^i_r\|^2}
\]
Now, since \( c^i_r \) is the mean of a subset of \( A_{t+1}^* \),
\[
\|c^i_r - c^{i-1}_r\|^2 \leq \phi^{i-1}_r
\]
Hence
\[
\sum_{r} n^*_r \|c^i_r - c^{i-1}_r\|^2 \leq n\phi^{i-1}_r
\]
On the other hand,
\[
\sum_{r} n^*_r \|c^i_r - c^{i-1}_r\|^2 \leq \sum_{r} n^*_r \|c^i_r - m(A_{t+1}^*)\|^2
\]
\[
\leq n \sum_{r} (\phi^{i-1}_r - \phi(m(A_{t+1}^*)))
\]
\[
= n(\phi^{i-1} - \phi(m(A_{t+1}^*))) \leq n(\phi^{i-1} - \phi^*)
\]
Now we first bound \( [11] \):
\[
\sum_{r} n^*_r (c^i_r - c^*_r, c^*_r - E[c^i_r | F_{i-1}])
\]
\[
\leq \sqrt{\Delta^{i-1} \sqrt{n} \hat{\phi}^{i-1} + \sqrt{\Delta^{i-1} \sqrt{n} \phi^{i-1} - \phi^*}}
\]
\[
\leq \sqrt{b_0 \phi^*} \sqrt{n(b_0 + 1) \phi^*} + \sqrt{nb_0 \phi^*} \leq 2(b_0 + 1)\sqrt{n} \phi^*
\]
To bound \( [12] \),
\[
\sum_{r} n^*_r \|c^i_r - c^*_r\|^2
\]
\[
\leq \sum_{r} n^*_r \|c^i_r - c^{i-1}_r\|^2 + \sum_{r} n^*_r \|c^{i-1}_r - c^*_r\|^2
\]
\[
\leq 2\sum_{r} n^*_r \|c^i_r - c^{i-1}_r\|^2 + 4\sum_{r} n^*_r \|c^{i-1}_r - c^*_r\|^2
\]
\[
\leq 2n(\phi^{i-1} + 2\Delta^{i-1}) \leq 2n(b_0 + 1)\phi^* + 4b_0 \phi^* \leq 4n(b_0 + 1)\phi^*
\]

**Claim 2.** In the context of Algorithm \( \#1 \), if \( \forall c^i \in C^t, c^i \in conv(X) \), then \( \forall c^{i+1} \in C^{i+1}, c^{i+1} \in conv(X) \).

**Proof of Claim.** By the update rule in Algorithm \( \#1 \), \( c^{i+1} \) is a convex combination of \( c^i \) and \( c^{i-1} \), where \( c^{i+1} \) is the mean of a subset of \( X \), and hence \( c^i \) is in \( conv(X) \). Since both \( c^i \) and \( c^{i+1} \) are in \( conv(X) \), \( c^{i+1} \in conv(X) \). \( \square \)

**Lemma 25 (technical lemma).** For any fixed \( b \in (1, 2] \). If \( C \geq b \frac{1}{\delta^2} \frac{1}{\delta^2} \delta \leq b \frac{1}{\delta^2} \), and \( t \geq (\frac{AC}{b-1} \ln \frac{1}{} \frac{1}{\delta} \frac{1}{\delta}) \), then \( t^{b-1} - 2C \ln t - C \ln \frac{b}{b} > 0 \).

**Proof.** Let \( f(t) := t^{b-1} - 2C \ln t - C \ln \frac{b}{b} \). Taking derivative, we get \( f'(t) = (b - 1)tb^{b-2} - \frac{2C}{t} \geq 0 \) when \( t \geq (\frac{2C}{b-1}) \). Since \( \ln \frac{1}{\delta^2} \frac{1}{\delta^2} \delta \leq \delta \), \( \delta \geq \frac{1}{\delta^2} \frac{1}{\delta^2} \delta \), \( \delta \geq (\frac{AC}{b-1} \ln \frac{1}{\delta} \frac{1}{\delta}) \), it suffices to show \( f((\frac{1}{\delta^2} \frac{1}{\delta^2} \delta)) > 0 \) for our statement to hold. \( f((\frac{1}{\delta^2} \frac{1}{\delta^2} \delta)) = \ln \frac{1}{\delta^2} \frac{1}{\delta^2} \delta - 2C \ln \ln \frac{1}{\delta^2} \frac{1}{\delta^2} \delta - C \ln \frac{b}{b} = \ln \frac{b}{b} \frac{1}{\delta^2} \frac{1}{\delta^2} \delta - 2C \ln \ln \frac{1}{\delta^2} \frac{1}{\delta^2} \delta - C \ln \frac{b}{b} = \frac{1}{\delta^2} \frac{1}{\delta^2} \delta \ln \frac{b}{b} - 2C \ln \ln \frac{1}{\delta^2} \frac{1}{\delta^2} \delta + C \ln \frac{b}{b} \left(\frac{1}{\delta^2} \frac{1}{\delta^2} \delta - 1 \right) > 0 \), where the first term is greater than zero because \( x - \ln(2x) > 0 \) for \( x > 0 \), and the second term is greater than zero by our assumption on \( C \). \( \square \)
Lemma 26 (Lemma D1 of [3]). Consider a nonnegative sequence \( \{u_t : t \geq t_0\} \), such that for some constants \( a, b > 0 \) and for all \( t > t_0 \geq 0 \), \( u_t \leq (1 - \frac{a}{t+1})u_{t-1} \). Then, if \( a > 1 \),
\[
\begin{align*}
\phi_{\text{batch}} &= \frac{b}{a - 1(1 + \frac{1}{t_0+1})^{a+1}} \frac{1}{(t+1)}
\end{align*}
\]

12 Appendix F: additional experiments

Our second set of experiments serves to corroborate our observations from the initial experiments, and to further explore the convergence behavior subject to different factors. To this end, we include two more benchmark datasets, \texttt{mnist} and \texttt{covtype}, a simulated dataset \texttt{gauss}, and add stochastic \( k \)-means with a constant learning rate. Instead of running the algorithm for only 100 iterations, we adopt a setup that is more akin to what is commonly used in practice — we divide the convergence into 20 epochs, where the epoch lengths are chosen to be one of 60, 600, and 6000 iterations.

The “burn-in” effect explained by a constant \( t_o \)

From our previous experiments, we observe that the initial phase of convergence is sometimes slower than \( \Theta(\frac{1}{t}) \) (e.g., in Figure 2a). This phenomenon also shows up, and in fact more frequently, when we turn to other datasets. Here is our explanation: the \( \frac{1}{t} \) (let \( b \) be some constant) model of convergence is not exactly what was derived from our theorems: the exact form of convergence rate in Theorem 1 and 2, which we hide behind the Big-O notation, is in fact \( \frac{1}{rt_o} \), where \( t_o \) is part of the learning rate parameter. After taking into account \( t_o \), our theoretical convergence rate well-matches our empirical observations. For example, in Figure 2 when \( t_o \) is set to be 60 or higher, the actual convergence can be simulated by (a proxy to) our theoretical bound \( \frac{c}{t_o} \). Note the practical requirement on \( t_o \) is much more optimistic than the lower bound in Theorem 1, i.e.,
\[
\frac{1}{t_o} \geq \frac{1}{r_{min}} \cdot \frac{1}{\frac{1}{n^2 \ln^2 n}}
\]

Again, we observe that the convergence rate of stochastic \( k \)-means is not sensitive to the choice of \( t_o \), despite the fact that the latter plays a role in explaining the convergence rate.

Runtime vs final \( k \)-means cost

Here, we compare the \( k \)-means cost achieved by stochastic \( k \)-means with different learning rates and epoch lengths to that achieved by batch \( k \)-means after 20 iterations. Each entry in the table is computed as \( \frac{C}{\text{batch}} \). \( C \) is the \( k \)-means cost of stochastic \( k \)-means after 20 iterations, with \( T = 20 \times E \), where \( E \) is a particular epoch length. \( \phi_{\text{batch}} \) is the final \( k \)-means cost of batch \( k \)-means. As shown in Table 1, the final \( k \)-means cost of stochastic \( k \)-means, using epoch length of 600, is already comparable to its batch counterpart. On the other hand, the data sizes of \texttt{mnist}, \texttt{covtype}, \texttt{gauss}, \texttt{rcv1} are 60k, 500k, 600k, and 800k, respectively. So even using the largest epoch length, 6k, stochastic \( k \)-means would save at least one-tenth of the computation in comparison to batch \( k \)-means. From the convergence plots (Figure 4 and 5), we see that the convergence behavior of stochastic \( k \)-means is not sensitive to the choice of learning rate. Here, we observe that learning rate does not affect the final \( k \)-means cost too much either; even a constant learning rate works!

Significance of different factors to convergence

Finally, we summarize the impact of different factors on the convergence behavior of stochastic \( k \)-means based on our experiments:

- Mini-batch size \( m \): the larger \( m \) is, the convergence becomes more stable and faster.
- Number of clusters \( k \): the smaller \( k \) is, the convergence becomes more stable and faster.
- Dataset: although \( \frac{k}{m} \) is observed for all datasets, stochastic \( k \)-means seems to favor certain datasets to others. For example, on \texttt{rcv1}, almost \( \frac{k}{T} \) (and sub-linear when \( m \) is larger) convergence rate is observed.
- Learning rate: the algorithm is not sensitive to the choice of learning rate.
Table 1: Final $k$-means cost relative to batch $k$-means: flat stands for our analyzed learning rate in \([6]\), and const for a fixed learning rate, which we set to be $\frac{1}{\sqrt{E}}$. For the flat learning rate, we arbitrarily choose $c' = 4$, and $t_0$ to be one of \(\{10, 60, 600, 6000\}\), which ever gives the lowest $k$-means cost.

<table>
<thead>
<tr>
<th></th>
<th>covtype</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$</td>
<td>$E=60$, flat, BBS, const</td>
<td>$E=600$, flat, BBS, const</td>
<td>$E=6k$, flat, BBS, const</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.93, 0.92, 0.93</td>
<td>0.99, 0.93, 0.99</td>
<td>1.03, 1.01, 1.03</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>1.13, 1.12, 1.13</td>
<td>1.02, 1.03, 1.02</td>
<td>1.01, 1.01, 1.01</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>1.15, 1.10, 1.15</td>
<td>1.05, 1.07, 1.05</td>
<td>1.02, 1.01, 1.02</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>mnist</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$</td>
<td>$E=60$, flat, BBS, const</td>
<td>$E=600$, flat, BBS, const</td>
<td>$E=6k$, flat, BBS, const</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1.07, 1.07, 1.07</td>
<td>1.02, 1.02, 1.02</td>
<td>1.03, 1.02, 1.03</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>1.15, 1.15, 1.15</td>
<td>1.06, 1.07, 1.06</td>
<td>1.02, 1.02, 1.02</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>1.18, 1.18, 1.18</td>
<td>1.07, 1.06, 1.07</td>
<td>1.02, 1.02, 1.02</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>rcv1</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$</td>
<td>$E=60$, flat, BBS, const</td>
<td>$E=600$, flat, BBS, const</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1.03, 1.03, 1.03</td>
<td>1.02, 1.02, 1.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>1.06, 1.06, 1.06</td>
<td>1.06, 1.06, 1.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>1.09, 1.09, 1.09</td>
<td>1.07, 1.07, 1.07</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>gauss</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$</td>
<td>$E=60$, flat, BBS, const</td>
<td>$E=600$, flat, BBS, const</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1.05, 1.07, 1.05</td>
<td>1.03, 1.03, 1.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>1.16, 1.14, 1.16</td>
<td>1.07, 1.05, 1.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>1.11, 1.11, 1.11</td>
<td>1.02, 1.02, 1.02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Convergence rate of stochastic $k$-means

Figure 4: Experiments on covtype
Figure 5: Experiments on \textit{mnist}