7 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^t 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 $1 \le k' \le k$. 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 $Cl(\cdot)$ 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 $\{A^*\}_{[k]} \subset \{A\}_{[k]}$ 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' clusters, 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' ($k' \ge k$) 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 1. Without loss of generality, we let $\pi(r) = r, \forall r \in [k]$. Let $\rho^r_{out} := \frac{|\cup_{s\neq r}(A_s \cap A^*_r)|}{n^*_r}$, and $\rho^r_{in} := \frac{|\cup_{s\neq r}(A_r \cap A^*_s)|}{n^*_r}$; let $\rho_{\max} := \max_r \frac{|A_r \triangle A^*_r|}{n^*_r}$.

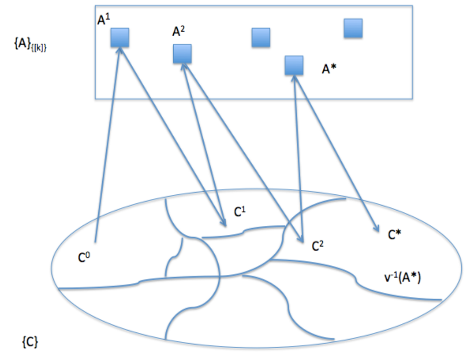

Figure 3: An illustration of one run of batch k-means in the solution spaces: the rectangle represent the enlarged space of clusterings $\{A\}_{[k]}$ and the ellipse represent the centroidal space $\{C\}$, which is partitioned into equivalences classes. The arrows represent k-means updates as mappings $v : \{C\} \to \{A\}_{[k]}$ and $m : \{A\}_{[k]} \to \{C\}$. The algorithm starts at C^0 and stops at C^* after three iterations, where $C^* = m(A^*) \in Cl(v^{-1}(A^*))$.

Clearly, $(\rho^r_{out} + \rho^r_{in}) = \frac{|A_r \triangle A^*_r|}{n^*_r} \le \rho_{\max}$, by our definition. Now, similar to [19], we can get $\|m(A_r) - c^*_r\| = \|\frac{(1-\rho^r_{out})n^*_r m(A_r\cap A^*_r) + \sum_{s\neq r}\sum_{x\in A_r\cap A^*_s} x}{(1-\rho^r_{out}+\rho^r_{in})n^*_r} - c^*_r\| \le \frac{1-\rho_{out}}{1-\rho^r_{out}+\rho^r_{in}}\|m(A_r \cap A^*_r) - c^*_r\| + \frac{\|\sum_{s\neq r}\sum_{x\in A_r\cap A^*_s} x - c^*_r\|}{(1-\rho^r_{out}+\rho^r_{in})n^*_r}$ And as in [19], we get $(1-\rho_{out})\|m(A_r \cap A^*_r) - c^*_r\| \le \frac{\sqrt{\rho^r_{out}\phi^*_r}}{\sqrt{n^*_r}}$. Now we bound the second term: by Cauchy-Schwarz inequality, $\|\sum_{s\neq r}\sum_{x\in A_r\cap A^*_s} x - c^*_r\|^2 \le (\sum_{s\neq r}\sum_{x\in A_r\cap A^*_s} 1^2)(\sum_{s\neq r}\sum_{x\in A_r\cap A^*_s}\|x - c^*_r\|^2) = \rho^r_{in} n^*_r \sum_{s\neq r}\sum_{x\in A_r\cap A^*_s}\|x - c^*_r\|^2$. Thus, $\forall r \in [k]$, $\|m(A_r) - c^*_r\|^2 \le 4\frac{\rho^r_{out}\phi^*_r}{n^*_r} + 4\frac{\rho^r_{in}\sum_{s\neq r}\sum_{x\in A_r\cap A^*_s}\|x-c^*_r\|^2}{n^*_r}$, where we use the assumption that $\rho_{\max} < \frac{1}{4} < 1 - \frac{1}{\sqrt{2}}$. Summing over all r, $\sum_r n^*_r\|m(A_r) - c^*_r\|^2 \le 4\rho_{\max}\sum_r(\phi^*_r + \sum_{s\neq r}\sum_{x\in A_r\cap A^*_s}\|x - c^*_r\|^2)$. By Lemma 3, $\sum_r\sum_{s\neq r}\sum_{x\in A_r\cap A^*_s}\|x - c^*_r\|^2$ can be upper bounded by $\phi(C') + \sum_r n_r\|m(A_r) - c^*_r\|^2 = \phi(C') + \sum_r(1 - \rho^r_{out} + \rho^r_{in})n^*_r\|m(A_r) - c^*_r\|^2 \le \phi(C') + (1+\rho_{\max})\sum_r n^*_r\|m(A_r) - c^*_r\|^2$. Substituting this into the previous inequality, we have $(1 - 4\rho_{\max}(1+\rho_{\max}))\sum_r n^*_r\|m(A_r) - c^*_r\|^2 \le 4\rho_{\max}(\phi^* + \phi(C'))$. Thus, $\sum_r n^*_r\|m(A_r) - c^*_r\|^2 \le \frac{\rho_{\max}}{1-4\rho_{\max}(1+\rho_{\max})}[\phi^* + \phi(C')]$. By our assumption, $\rho_{\max} \le \frac{b}{5b+4(1+\frac{\phi(C)}{\phi^*})} < \frac{1}{4}$, so $\frac{\rho_{\max}}{1-4\rho_{\max}(1+\rho_{\max})} \le \frac{\rho_{\max}}{1-5\rho_{\max}} \le \frac{b}{1+\frac{\phi(C)}{\phi^*}}$, and $\frac{\rho_{\max}}{1-4\rho_{\max}(1+\rho_{\max})}[\phi^* + \phi(C')] \le b\phi^*$, since $\phi(C') \le \phi(C)$ (equality holds if C is a stationary point). $\square$

Lemma 3. *Fix any target clustering C^*, and another clustering C with a matching $\pi : [k] \to [k]$. Let $C' :=$*

$\{m(A_r), r \in [k]\}$. *Then*

$$\sum_r \sum_{s\neq r} \sum_{x\in A_{\pi(r)}\cap A_s^*} \|x - c_r^*\|^2$$
$$\leq \phi(C') - \sum_r \phi(c_r^*; A_{\pi(r)} \cap A_r^*) + \sum_r n_r \|m(A_r) - c_r^*\|^2$$

Proof. Without loss of generality, we let $\pi(r) = r$.

$$\phi(C') - \phi(C^*) = \sum_r \sum_{x\in A_r} \|x - c_r^*\|^2 - \sum_r \sum_{x\in A_r^*} \|x - c_r^*\|^2$$
$$+ \sum_r \sum_{x\in A_r} \|x - m(A_r)\|^2 - \sum_r \sum_{x\in A_r} \|x - c_r^*\|^2$$

So $\sum_r \sum_{x\in A_r} \|x-c_r^*\|^2 - \sum_r \sum_{x\in A_r^*} \|x-c_r^*\|^2 = \phi(C') - \phi(C^*) - \sum_r \sum_{x\in A_r} \|x - m(A_r)\|^2 + \sum_r \sum_{x\in A_r} \|x - c_r^*\|^2 \leq \phi(C) - \phi(C^*) + \sum_r n_r \|m(A_r) - c_r^*\|^2$. Now, we claim $\sum_r \sum_{x\in A_r} \|x-c_r^*\|^2 - \sum_r \sum_{x\in A_r^*} \|x-c_r^*\|^2 = \sum_r \sum_{s\neq r} \sum_{x\in A_r\cap A_s^*} \{\|x-c_r^*\|^2 - \|x-c_s^*\|^2\}$. This is because we can enumerate x using clustering $\cup_r A_r$: 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 $x \in A_r \cap A_s^*$ for some $s \neq r$, which means the difference is $\|x - c_r^*\|^2 - \|x - c_s^*\|^2$ (and this term is positive by optimality of clustering $\cup_r A_r^*$ fixing $\{c_r^*\}$). Thus, $\sum_r \sum_{s\neq r} \sum_{x\in A_r\cap A_s^*} \|x - c_r^*\|^2 = \sum_r \sum_{x\in A_r} \|x-c_r^*\|^2 - \sum_r \sum_{x\in A_r^*} \|x-c_r^*\|^2 + \sum_r \sum_{s\neq r} \sum_{x\in A_r\cap A_s^*} \|x - c_s^*\|^2 \leq \phi(C') - \phi(C^*) + \sum_r n_r \|m(A_r) - c_r^*\|^2 + \sum_r \sum_{s\neq r} \sum_{x\in A_r\cap A_s^*} \|x - c_s^*\|^2 = \phi(C') - \sum_r \phi(c_r^*; A_r \cap A_r^*) + \sum_r n_r \|m(A_r) - c_r^*\|^2$, where the last equality is by observing that $\phi(C^*) = \sum_r \sum_{A_r\cap A_r^*} \|x - c_r^*\|^2 + \sum_r \sum_{s\neq r} \sum_{x\in A_r\cap A_s^*} \|x-c_s^*\|^2$. □

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 $\|x - c_r\| = \|x - c_s\|$ if and only if $\|\bar{x} - c_r\| - \|\bar{x} - c_s\|$. "$2 \implies 3$": let $A \in V(C) \cap X$ be the clustering achieving the zero margin, and consider $x \in A_r \cup A_s$ s.t. $\|\bar{x} - c_r\| - \|\bar{x} - c_s\|$; without loss of generality, assume $x \in A_r$ 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_s. Then $A' \in V(C) \cap X$ and $|V(C) \cap X| \geq 2 > 1$. "$3 \implies 1$": Suppose otherwise. Then every point x has a unique center that minimizes its distance to it, which means the clustering determined by $V(C) \cap A$ is unique. A contradiction. □

Lemma 5. *If $C^* \in \{C^*\}$, 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 $v \circ m(A) = v(C^*) = A$. 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:*

$$\text{For } C' \text{ s.t. } \Delta(\cdot,\cdot) \text{ is defined },$$
$$\Delta(C', C) < \delta \implies C' \in v^{-1}(A) \tag{9}$$

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. $\forall x \in A_r$, $\forall r \in [k], s \neq r$,

$$\|x - c_r\| < \|x - c_s\| - 2\sqrt{\delta}$$

Let π^* be a permutation such that $\Delta(C', C)$ is defined. We have $\forall x \in A_r, r \in [k], s \neq r$,

$$\|x - c'_{\pi^*(s)}\| - \|x - c'_{\pi^*(r)}\| \geq \|x - c_s\| - \|c'_{\pi^*(s)} - c_s\|$$
$$-(\|x - c_r\| + \|c_r - c'_{\pi^*(r)}\|) > \|x - c_s\| - \|x - c_r\| - 2\sqrt{\delta} \geq 0$$

where the second inequality is by the fact that

$$\max_r \|c'_{\pi^*(r)} - c_r\|^2 \leq \Delta(C', C) < \delta$$

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 δ 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 bisector. 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 $\exists r_{\min} > 0$ such that $\forall C^* \in \{C^*\}_{[k]}$, C^* is a $(r_{\min}, 0)$-stable stationary point.*

Proof. Fix any k 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 $\exists r^* > 0$, such that the following statement holds:

$$\Delta(C, C^*) < r^* \phi^* \implies C \in v^{-1}(A^*)$$

Since C^* is a non-boundary point, there is a permutation π_o of $[k]$ such that $\forall x \in A_r, \forall r \in [k]$ and $\forall s \neq r$,

$$\|x - c^*_{\pi_o(r)}\| < \|x - c^*_{\pi_o(s)}\|$$

We choose $r^* > 0$ so that $\forall x \in A_r, \forall r \in [k], \forall s \neq r$,

$$\|x - c^*_{\pi_o(r)}\| \leq \|x - c^*_{\pi_o(s)}\| - 2\sqrt{r^*\phi^*},\ \forall r \in [k], s \neq r$$

with equality holds for at least one triple of (x, r, s). Let π^* be a permutation satisfying

$$\pi^* = \arg\min_{\pi} \sum_{r\in[k]} n^*_r \|c_{\pi(r)} - c^*_r\|^2$$

Let $\pi' := \pi^* \circ \pi_o$. We have $\forall (x, r, s)$ triples,

$$\begin{aligned}&\|x - c_{\pi'(s)}\| - \|x - c_{\pi'(r)}\| \\ &\geq \|x - c^*_{\pi_o(s)}\| - \|c^*_{\pi_o(s)} - c_{\pi'(s)}\| \\ &-(\|x - c^*_{\pi_o(r)}\| + \|c^*_{\pi_o(r)} - c_{\pi'(r)}\|) \\ &> \|x - c^*_{\pi_o(s)}\| - \|x - c^*_{\pi_o(r)}\| - 2\sqrt{r^*\phi^*} \geq 0\end{aligned}$$

where the second inequality is by the fact that

$$\begin{aligned}&\max_r \|c_{\pi^*(r)} - c^*_r\|^2 \leq \Delta(C, C^*) < r^*\phi^* \\ &\implies \max_r \|c_{\pi^*(r)} - c^*_r\| < \sqrt{r^*\phi^*}\end{aligned}$$

Since π' 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 $C \in v^{-1}(A^*)$. Since by our definition, r^* is unique for each C^*. Since $\{C^*\}_{[k]}$ is finite, taking the minimum over all such r^*, i.e., $r_{\min} := \min_{C^*\in\{C^*\}_{[k]}} r^*$ 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 $\exists r_{\min} > 0$ s.t.*

1. *$\forall C^* \in \{C^*\}_{[k]}$, C^* is a $(r_{\min}, 0)$-stable stationary point.*
2. *Let $m(A') \notin \{C^*\}_{[k]}$ for some $A' \in \{A\}$ and let $C' \in Cl(v^{-1}(A'))$, then $\Delta(C', m(A)) \geq r_{\min}\phi(m(A))$.*

Proof. By Lemma 8, $\exists r^*_{\min} > 0$ s.t. $\forall C^*$, C^* is $r^*_{\min}$-stable. Furthermore, by Lemma 7, $\exists r'_{\min} > 0$ 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 1. For all $r \in [k]$,

$$n^*_r \|c_r - c^*_r\|^2 \leq \Delta(C, C^*) \leq b\phi^*$$

so $\|c_r - c^*_r\| \leq \sqrt{\frac{b\phi^*}{n^*_r}}$. Then for all $r \neq s$,

$$\begin{aligned}&\|c_r - c^*_r\| + \|c_s - c^*_s\| \leq \sqrt{b}\sqrt{\phi^*}(\frac{1}{\sqrt{n^*_r}} + \frac{1}{\sqrt{n^*_s}}) \\ &= \frac{\sqrt{b}}{f} f\sqrt{\phi^*}(\frac{1}{\sqrt{n^*_r}} + \frac{1}{\sqrt{n^*_s}}) \leq \frac{\sqrt{b}}{f}\Delta_{rs} \leq \frac{1}{16}\Delta_{rs}\end{aligned}$$

where the second inequality is by (B), and the last inequality by our assumption on b. Thus, we may apply Lemma 17 to get $\frac{|A_r \triangle A^*_r|}{n^*_r} \leq \frac{b}{f^3}$ for all r, proving the first statement. Now by Lemma 18, $\phi(C) \leq (b+1)\phi^*$, so

$$\begin{aligned}\frac{\alpha b}{5\alpha b + 4(1 + \frac{\phi(C)}{\phi^*})} &\geq \frac{\alpha b}{5\alpha b + 4(2+b)} \\ &\geq \frac{\alpha b}{5\alpha f^2/16^2 + 4(2 + f^2/16^2)} \\ &\geq \frac{b}{f^3(\alpha)} \geq \frac{|A_r \triangle A^*_r|}{n^*_r}\end{aligned}$$

where the third inequality holds since $f \geq \max\{64^2, \frac{5\alpha+5}{16^2\alpha}\}$ by (B). This proves the second statement since C^* is then $(\frac{f^2}{16^2}, \alpha)$-stable by Definition 4. □

9 Appendix C: Proof of Theorem 3

Theorem 3. *Fix any $0 < \delta \leq \frac{1}{e}$. Suppose C^* is (b_o, α)-stable. If we run Algorithm 1 with parameters satisfying*

$$m > \frac{\ln(1 - \sqrt{\alpha})}{\ln(1 - \frac{4}{5}p^*_{\min})}$$

$$c' > \frac{\beta}{2[1 - \sqrt{\alpha} - (1 - \frac{4}{5}p^*_{\min})^m]} \text{ with } \beta \geq 2$$

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

Then if at some iteration i, $\Delta^i \leq \frac{1}{2}b_o\phi^*$, we have $\forall t > i$,

$$Pr(\Omega_t) \geq 1 - \delta \quad \textit{and}$$

$$\begin{aligned}&E_t[\Delta^t] \leq (\frac{t_o + i + 1}{t_o + t + 1})^\beta \Delta^i \\ &+ \frac{(c')^2 B}{\beta - 1}(\frac{t_o + i + 2}{t_o + i + 1})^{\beta+1}\frac{1}{t_o + t + 1}\end{aligned}$$

where $B := 4(b_o + 1)n\phi^*$.

9.1 Proofs leading to Theorem 3

In the subsequent analysis, we let

$$\beta^t := 2c' \min_r p^t_r(m)(1 - \frac{\max_r p^t_r(m)}{\min_s p^t_s(m)}\sqrt{\alpha})$$

where

$$\begin{aligned}&p^t_r(m) := Pr\{c^{t-1}_r \text{ is updated at } t \text{ with sample size } m\} \\ &= 1 - (1 - \frac{n^{t-1}_r}{n})^m\end{aligned}$$

So,

$$\beta^t = 2c'(\min_r p^t_r(m) - \sqrt{\alpha}\max_s p^t_s(m))$$

The noise terms appearing in our analysis are:

$$E[\sum_r \sum_{x\in A^{t+1}_r} \|x - \hat{c}^{t+1}_r\|^2 + \phi^t | F_t] \tag{10}$$

$$\sum_r n^*_r \langle c^{t-1}_r - c^*_r, \hat{c}^t_r - E[\hat{c}^t_r | F_{t-1}]\rangle \tag{11}$$

$$\sum_r n^*_r \|\hat{c}^t_r - c^*_r\|^2 \tag{12}$$

In the analysis of this section, we use $E_t[\cdot]$ as a shorthand notation for $E[\cdot|\Omega_t]$, where Ω_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 with the algorithm's parameters m, c', and t_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 $\phi^*|\Omega_t$
- $Pr(\Omega_t \setminus \Omega_{t+1})$ is negligible $|\Omega_t, \beta^t \geq 2$, bounded noise
- $E_t[\Delta^t|F_{t-1}] \leq (1 - \frac{\beta^t}{t_o+t})\Delta^{t-1} + \epsilon^t \ |\Omega_t$

  where ϵ^t, the noise term, decreases of order $O(\frac{1}{t^2})$.

Lemma 10. *Suppose C^* is (b_o, α)-stable. If*

$$m > \frac{\ln(1-\sqrt{\alpha})}{\ln(1-\frac{4}{5}p^*_{\min})}$$

and

$$c' > \frac{\beta}{2[1-\sqrt{\alpha}-(1-\frac{4}{5}p^*_{\min})^m]}$$

Then conditioning on Ω_t, we have $\beta^t \geq \beta$.

Proof. Let's first consider $p^t_r(1) = \frac{n^{t-1}_r}{n}$. Conditioning on Ω_t, using the fact that C^* is (b_o, α)-stable, we have

$$\frac{n^{t-1}_r}{n} \geq p^*_{\min}(1 - \max_r \frac{|A^t_r \triangle A^*_r|}{n^*_r})$$
$$\geq p^*_{\min}(1 - \frac{\alpha b_o}{5\alpha b_o + 4(1+\frac{\phi^t}{\phi^*})}) \geq \frac{4}{5}p^*_{\min}$$

And hence,

$$\min_r p^t_r(m) \geq 1 - (1-\frac{4}{5}p^*_{\min})^m$$

Now,

$$\beta^t \geq 2c'(\min_r p^t_r(m) - \sqrt{\alpha})$$
$$\geq 2c'(1-(1-\frac{4}{5}p^*_{\min})^m - \sqrt{\alpha}) \geq \beta$$

where the last inequality is by our requirement on c' and the fact that $1-(1-\frac{4}{5}p^*_{\min})^m - \sqrt{\alpha} > 0$ by our requirement on m. ☐

Lemma 11. *Suppose C^* is (b_o, α)-stable. Then if we apply one step of Algorithm 1, with m, c' satisfying conditions in Lemma 10, then conditioning on Ω_i,*

$$\Delta^i \leq \Delta^{i-1}(1-\frac{\beta}{t_o+i}) + [\frac{c'}{t_o+i}]^2 \sum_r n^*_r \|\hat{c}^i_r - c^*_r\|^2$$
$$+\frac{2c'}{t_o+i}\sum_r n^*_r \langle c^{i-1}_r - c^*_r, \xi^i_r\rangle$$

where $\xi^i_r := \hat{c}^i_r - E[\hat{c}^i_r|F_{i-1}]$.

Proof. Let $\Delta^i_r := n^*_r\|c^i_r - c^*_r\|^2$, so $\Delta^i = \sum_r \Delta^i_r$, and we use p^t_r as a shorthand for $p^t_r(m)$. By the update rule of Algorithm 1,

$$\Delta^i_r = n^*_r\|(1-\eta^i)(c^{i-1}_r - c^*_r) + \eta^i(\hat{c}^i_r - c^*_r)\|^2$$
$$\leq n^*_r\{(1-2\eta^i)\|c^{i-1}_r - c^*_r\|^2 + 2\eta^i\langle c^{i-1}_r - c^*_r, \hat{c}^i_r - c^*_r\rangle$$
$$+(\eta^i)^2[\|c^{i-1}_r - c^*_r\|^2 + \|\hat{c}^i_r - c^*_r\|^2]\}$$

Let $\xi^i_r = \hat{c}^i_r - E[\hat{c}^i_r|F_{i-1}]$, where

$$E[\hat{c}^i_r|F_{i-1}] = (1-p^i_r)c^{i-1}_r + p^i_r m(A^i_r)$$

Since

$$\langle c^{i-1}_r - c^*_r, \hat{c}^i_r - c^*_r\rangle = \langle c^{i-1}_r - c^*_r, E[\hat{c}^i_r|F_{i-1}] + \xi^i_r - c^*_r\rangle$$
$$\leq (1-p^i_r)\|c^{i-1}_r - c^*_r\|^2$$
$$+p^i_r\|m(A^i_r) - c^*_r\|\|c^{i-1}_r - c^*_r\| + \langle c^{i-1}_r - c^*_r, \xi^i_r\rangle$$

We have

$$\Delta^i_r \leq n^*_r\{-2\eta^i[\|c^{i-1}_r - c^*_r\|^2 - (1-p^i_r)\|c^{i-1}_r - c^*_r\|^2$$
$$-p^i_r\|c^{i-1}_r - c^*_r\|\|m(A^i_r) - c^*_r\|] + \|c^{i-1}_r - c^*_r\|^2$$
$$+2\eta^i\langle \xi^i_r, c^{i-1}_r - c^*_r\rangle + (\eta^i)^2[\|c^{i-1}_r - c^*_r\|^2 + \|\hat{c}^i_r - c^*_r\|^2]\}$$
$$\leq n^*_r\{-\frac{2c'}{t_o+i}\min_r p^t_r\|c^{i-1}_r - c^*_r\|^2$$
$$+\frac{2c'}{t_o+i}\max_s p^t_s\|c^{i-1}_r - c^*_r\|\|m(A^i_r) - c^*_r\|$$
$$+\|c^{i-1}_r - c^*_r\|^2 + 2\eta^i\langle \xi^i_r, c^{i-1}_r - c^*_r\rangle$$
$$+(\eta^i)^2[\|c^{i-1}_r - c^*_r\|^2 + \|\hat{c}^i_r - c^*_r\|^2]\}$$

Note

$$\sum_r n^*_r\|c^i_r - c^*_r\|\|m(A^i_r) - c^*_r\|$$
$$\leq \sqrt{(\sum_r n^*_r\|c^{i-1}_r - c^*_r\|^2)(\sum_r n^*_r\|m(A^i_r) - c^*_r\|^2)}$$
$$= \sqrt{\Delta^{i-1}\Delta(m(A^i), C^*)} \leq \sqrt{\alpha}\Delta^{i-1}$$

where the first inequality is by Cauchy-Schwartz and the last inequality is by applying Lemma 1. Finally, summing over Δ^i_r, we get

$$\Delta^i = \sum_r \Delta^i_r \leq \Delta^{i-1}[1 - \frac{2c'}{t_o+i}\min_r p^t_r(1 - \frac{\max_s p^t_s}{\min_r p^t_r}\sqrt{\alpha})]$$
$$+[\frac{c'}{(t_o+i)}]^2\sum_r n^*_r\|\hat{c}^i_r - c^*_r\|^2$$
$$+\frac{2c'}{(t_o+i)p^i_r}\sum_r n^*_r\langle c^{i-1}_r - c^*_r, \xi^i_r\rangle$$
$$\leq \Delta^{i-1}(1-\frac{\beta}{t_o+i}) + [\frac{c'}{t_o+i}]^2\sum_r n^*_r\|\hat{c}^i_r - c^*_r\|^2$$
$$+\frac{2c'}{t_o+i}\sum_r n^*_r\langle c^{i-1}_r - c^*_r, \xi^i_r\rangle$$

The second inequality is by $\beta^t \geq \beta$, as proven in Lemma 10. ☐

Lemma 12. *Suppose X satisfies (A1), $C^o \in conv(X)$, and C^* is (b_o, α)-stable. If we run one step of Algorithm 1, with m, c' satisfying conditions in Lemma 10, then conditioning on Ω_i, we have, for any $\lambda > 0$,*

$$E_i\{\exp\{\lambda\Delta^i\}|F_{i-1}\}$$
$$\leq \exp\left\{\lambda\{(1-\frac{\beta}{t_0+i})\Delta^{i-1}+\frac{(c')^2B}{(t_0+i)^2}+\frac{\lambda(c')^2B^2}{2(t_0+i)^2}\}\right\}$$

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

$$E_i\{\exp(\lambda\Delta^i)|F_{i-1}\} \leq \exp\lambda[\Delta^{i-1}(1-\frac{\beta}{t_o+i})+\frac{(c')^2B}{(t_o+i)^2}]$$
$$E_i\{\exp\lambda\frac{2c'}{t_o+i}\sum_r n_r^*\langle c_r^{i-1}-c_r^*, \xi_r^i\rangle|F_{i-1}\}$$

Since

$$\frac{2\lambda c'}{i+t_0}\sum_r n_r^*\langle\xi_r^i, c_r^{i-1}-c_r^*\rangle \leq \frac{2\lambda c'}{i+t_0}B$$

and $E_i\{\frac{2\lambda c'}{i+t_0}\sum_r n_r^*\langle\xi_r^i, c_r^{i-1}-c_r^*\rangle|F_{i-1}\}=0$, by Hoeffding's lemma

$$E_i\left\{\exp\{\frac{2\lambda c'}{i+t_0}\sum_r n_r^*\langle\xi_r^i, c_r^{i-1}-c_r^*\rangle|F_{i-1}\}\right\}$$
$$\leq \exp\{\frac{\lambda^2(c')^2B^2}{2(i+t_0)^2}\}$$

Combining this with the previous bound completes the proof. □

Lemma 13 (adapted from [4])**.** *For any $\lambda > 0$, $E_i\{e^{\lambda\Delta^{i-1}}\} \leq E_{i-1}\{e^{\lambda\Delta^{i-1}}\}$*

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 $\omega' \in \Omega_{i-1}\setminus\Omega_i$, $\Delta^{i-1}(\omega) \leq b_o\phi^* < \Delta^{i-1}(\omega')$. Taking expectation over Ω_i and Ω_{i-1}, we get $E_i\{e^{\lambda\Delta^{i-1}}\} \leq E_{i-1}\{e^{\lambda\Delta^{i-1}}\}$. □

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

$$m > \frac{\ln(1-\sqrt{\alpha})}{\ln(1-\frac{4}{5}p^*_{\min})}$$

$$c' > \frac{\beta}{2[1-\sqrt{\alpha}-(1-\frac{4}{5}p^*_{\min})^m]} \text{ with } \beta \geq 2$$

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

Then

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

(here we used Δ^0 instead of Δ^i and treat the starting time, the i-th iteration in Theorem 3 as the zeroth iteration for cleaner presentation).

Proof. By Lemma 12, for any $\lambda > 0$,

$$E_i\{e^{\lambda\Delta^i}\} \leq E_i\{e^{\lambda\{(1-\frac{\beta}{t_o+i})\Delta^{i-1}}\}\exp\{\frac{\lambda(c')^2B}{(t_o+i)^2}+\frac{\lambda^2(c')^2B^2}{2(t_o+i)^2}\}$$
$$\leq E_{i-1}\{e^{\lambda^{(1)}\Delta^{i-1}}\}\exp\{\frac{\lambda(c')^2B}{(t_o+i)^2}+\frac{\lambda^2(c')^2B^2}{2(t_o+i)^2}\}$$

where $\lambda^{(1)} = \lambda(1-\frac{\beta}{t_o+i})$, and the second inequality is by Lemma 13. Similarly, the following recurrence relation holds for $k = 0, \ldots, i$:

$$E_{i-k}\{e^{\lambda^{(k)}\Delta^{i-k}}\} \leq E_{i-(k+1)}\{e^{\lambda^{(k+1)}\Delta^{i-k-1}}\}$$
$$\exp\{\frac{\lambda^{(k)}(c')^2B}{(t_o+i-k)^2}+\frac{(\lambda^{(k)})^2(c')^2B^2}{2(t_o+i-k)^2}\}$$

where $\lambda^{(0)} := \lambda$, and for $k \geq 1$, $\lambda^{(k)} := \Pi_{t=1}^k(1-\frac{\beta}{t_o+(i-t+1)})\lambda^{(0)}$.

Note (see, e.g., [4]) $\forall\beta > 0, k \geq 1$,

$$\lambda^{(k)} = \Pi_{t=1}^k(1-\frac{\beta}{t_o+(i-t+1)}) \leq (\frac{t_o+i-k+1}{t_o+i})^\beta$$

Since the bound is shrinking as β increases and $\beta \geq 2$,

$$\frac{\lambda^{(k)}}{(t_0+i-k)^2} \leq (\frac{t_o+i-k+1}{t_o+i})^2\frac{\lambda}{(t_o+i-k)^2} \leq \frac{4\lambda}{(t_o+i)^2}$$

Repeatedly applying the relation, we get

$$E_i\{e^{\lambda\Delta^i}\} \leq e^{\lambda^{(i)}\Delta^0}\exp\{\sum_{k=0}^{i-1}(\frac{4\lambda(c')^2B}{(t_o+i)^2}+\frac{4\lambda^2(c')^2B^2}{2(t_o+i)^2})\}$$
$$\leq \exp\{\lambda(\frac{t_o+1}{t_o+i})^\beta\Delta^0+[\lambda(c')^2B+\frac{\lambda^2(c')^2B^2}{2}]\frac{4i}{(t_o+i)^2}\}$$
$$\leq \exp\{\lambda(\frac{t_o+1}{t_o+i})^\beta\frac{1}{2}b_o\phi^*+[\lambda(c')^2B+\frac{\lambda^2(c')^2B^2}{2}]\frac{4i}{(t_o+i)^2}\}$$

Then we can apply the conditional Markov's inequality, for any $\lambda_i > 0$,

$$Pr(\omega \in \Omega_i\setminus\Omega_{i+1}) = Pr(\Delta^i > b_o\phi^*|\Omega_i)$$
$$= Pr(e^{\lambda_i\Delta^i} > e^{\lambda_i b_o\phi^*}|\Omega_i) \leq \frac{E[e^{\lambda_i\Delta_r^i}|\Omega_i]}{e^{\lambda_i b_o\phi^*}}$$

Combining this with the upper bound on $E_ie^{\lambda_i\Delta^i}$, we get

$$Pr(\omega \in \Omega_i\setminus\Omega_{i+1})$$
$$\leq \exp\{-\lambda_i\{\frac{1}{2}b_o[2-(\frac{t_o+1}{t_o+i})^\beta]$$
$$-(B+\frac{\lambda_iB^2}{2})\frac{4(c')^2i}{(t_o+i)^2}\}\}$$
$$\leq \exp\left\{-\lambda_i\{\frac{b_o\phi^*}{2}-(B+\frac{\lambda_iB^2}{2})\frac{4(c')^2i}{(t_o+i)^2}\}\right\}$$

since $i \geq 1$. We choose $\lambda_i = \frac{1}{\Delta}\ln\frac{(i+1)^2}{\delta}$ with $\Delta = \frac{b_o\phi^*}{4}$, and show that $\frac{b_o\phi^*}{2}-(B+\frac{\lambda_iB^2}{2})\frac{4(c')^2i}{(t_o+i)^2}$ is lower bounded by Δ.

Case 1: $B > \frac{\lambda_i B^2}{2}$. We get

$$\frac{1}{2}b_o\phi^* - (B + \frac{\lambda_i B^2}{2})\frac{4(c')^2 i}{(t_o+i)^2} \geq \Delta$$

since $t_o \geq \frac{128(c')^2(b_o+1)n}{b_o} = \frac{64(c')^2(b_o+1)n\phi^*}{\frac{1}{2}b_o\phi^*} = \frac{16(c')^2 B}{\frac{1}{2}b_o\phi^*}$.

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

$$\begin{aligned}\frac{1}{2}b_o\phi^* - (B + \frac{\lambda_i B^2}{2})\frac{4(c')^2 i}{(t_o+i)^2} \\ \geq 2\Delta - \lambda_i B^2\frac{4(c')^2 i}{(t_o+i)^2} \\ = 2\Delta - \frac{1}{\Delta}\ln\frac{(1+i)^2}{\delta}\frac{4(c')^2B^2 i}{(t_o+i)^2} \\ \geq 2\Delta - \frac{1}{\Delta}\ln\frac{(t_o+i)^2}{\delta}\frac{4(c')^2B^2(t_o+i)}{(t_o+i)^2}\end{aligned}$$

Now we show

$$\frac{1}{\Delta}\ln\frac{(t_o+i)^2}{\delta}\frac{4(c')^2B^2}{t_o+i} \leq \Delta$$

Since

$$\begin{aligned}t_o + i \geq t_o \geq 768(c')^2(1+\frac{1}{b_o})^2 n^2 \ln^2\frac{1}{\delta} \\ = \frac{48(c')^2B^2}{(\frac{1}{2}b_o\phi^*)^2}\ln^2\frac{1}{\delta}\end{aligned}$$

$\ln\frac{1}{\delta} \geq 1$, and $\frac{16(c')^2B^2}{(\frac{1}{2}b_o\phi^*)^2} \geq \frac{1}{3}$, we can apply Lemma 25 with $b = 2$, $C := \frac{16(c')^2B^2}{(\frac{1}{2}b_o\phi^*)^2}$, $t := t_o + i \geq (\frac{3C}{b-1}\ln\frac{1}{\delta})^{\frac{2}{b-1}}$, and get

$$\frac{4(c')^2B^2}{\Delta^2}\ln\frac{(t_o+i)^2}{\delta} := 2C\ln t + C\ln\frac{1}{\delta} < t^{b-1} = t_o + i$$

That is, $\frac{1}{\Delta}\ln\frac{(t_o+i)^2}{\delta}\frac{4(c')^2B^2}{t_o+i} \leq \Delta$. Thus, for both cases,

$$2\Delta - (B + \frac{\lambda_i B^2}{2})\frac{4(c')^2 i}{(t_o+i)^2} \geq= \Delta$$

and hence,

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

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

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

This proves the first statement. Taking expectation over Ω_t conditioning on filtration F_{t-1} with respect to the inequality derived in Lemma 11, we get

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

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

$$\begin{aligned}E_t[\Delta^t] &\leq E_t[\Delta^{t-1}](1 - \frac{\beta}{t_o+t}) + \frac{(c')^2B}{(t+t_o)^2} \\ &\leq E_{t-1}[\Delta^{t-1}](1 - \frac{\beta}{t_o+t}) + \frac{(c')^2B}{(t+t_o)^2}\end{aligned}$$

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

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

□

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 η_r^t as well as the update rule:

$$c_r^t \leftarrow (1-\eta_r^t)c_r^{t-1} + \eta_r^t\hat{c}_r^t$$

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 $\hat{c}_r^t = c_r^{t-1}$ and 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 η_r^t, and differentiates between a sampled and not-sampled cluster only through the definition of $\hat{c}_r^t$.

Proof leading to Theorem 1

Lemma 14. *Suppose* $\forall r \in [k]$, $\eta_r^t \leq \eta_{\max}^t$ *w.p. 1. Then,* $E[\phi^{t+1} - \phi^t|F_t] \leq -2\min_{r,t;p_r^{t+1}>0}\eta_r^{t+1}p_r^{t+1}(\phi^t - \tilde{\phi}^t) + (\eta_{\max}^{t+1})^2 6\phi^t$, *where* $\tilde{\phi}^t := \sum_r\sum_{x\in A_r^{t+1}}\|x - m(A_r^{t+1})\|^2$.

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

$$\begin{aligned}E_t[\phi^{t+1}] = E_t[\sum_{r=1}^k\sum_{x\in A_r^{t+2}}\|x - c_r^{t+1}\|^2] \\ \leq E_t[\sum_r\sum_{x\in A_r^{t+1}}\|x - c_r^{t+1}\|^2] \\ = E_t[\sum_r\sum_{x\in A_r^{t+1}}\|x - (1-\eta_r^{t+1})c_r^t - \eta_r^{t+1}\hat{c}_r^{t+1}\|^2] \\ = E_t[\sum_r\sum_{x\in A_r^{t+1}}(1-\eta_r^{t+1})^2\|x - c_r^t\|^2 \\ +(\eta_r^{t+1})^2\|x - \hat{c}_r^{t+1}\|^2 + 2\eta_r^{t+1}(1-\eta_r^{t+1})\langle x - c_r^t, x - \hat{c}_r^{t+1}\rangle]\end{aligned}$$

where the inequality is due to the optimality of clustering A^{t+2} for centroids C^{t+1}. Since

$$E_t[\hat{c}_r^{t+1}] = (1-p_r^{t+1})c_r^t + p_r^{t+1}m(A_r^{t+1})$$

we have

$$\begin{aligned}&\langle x - c_r^t, x - \hat{c}_r^{t+1}\rangle \\ &= (1-p_r^{t+1})\|x-c_r^t\|^2 + p_r^{t+1}\langle x - c_r^t, x - m(A_r^{t+1})\rangle\end{aligned}$$

Plug this into the previous inequality, we get

$$\begin{aligned}E_t[\phi^{t+1}] \le \sum_r (1-2\eta_r^{t+1})\phi_r^t + (\eta_r^{t+1})^2\phi_r^t \\ +(\eta_r^{t+1})^2 \sum_{x\in A_r^{t+1}} \|x-\hat{c}_r^{t+1}\|^2 \\ +2\eta_r^{t+1}\{(1-p_r^{t+1})\sum_{x\in A_r^{t+1}}\|x-c_r^t\|^2 \\ +p_r^{t+1}\sum_{x\in A_r^{t+1}}\langle x-c_r^t, x-m(A_r^{t+1})\rangle\} \\ = \phi^t - 2\sum_r \eta_r^{t+1}p_r^{t+1}\phi_r^t \\ +2\sum_r \eta_r^{t+1}p_r^{t+1}\sum_{x\in A_r^{t+1}}\langle x-c_r^t, x-m(A_r^{t+1})\rangle\} \\ +(\eta_r^{t+1})^2\phi_r^t + (\eta_r^{t+1})^2\sum_{x\in A_r^{t+1}}\|x-\hat{c}_r^{t+1}\|^2\end{aligned}$$

Now,

$$\begin{aligned}&\sum_{x\in A_r^{t+1}}\langle x-c_r^t, x-m(A_r^{t+1})\rangle \\ &= \sum_{x\in A_r^{t+1}}\langle x-m(A_r^{t+1})+m(A_r^{t+1})-c_r^t, x-m(A_r^{t+1})\rangle \\ &= \sum_{x\in A_r^{t+1}}\|x-m(A_r^{t+1})\|^2 \\ &+ \sum_{x\in A_r^{t+1}}\langle m(A_r^{t+1})-c_r^t, x-m(A_r^{t+1})\rangle = \phi_r^t\end{aligned}$$

since $\sum_{x\in A_r^{t+1}}\langle m(A_r^{t+1})-c_r^t, x-m(A_r^{t+1})\rangle = 0$, by property of the mean of a cluster. Then

$$\begin{aligned}E_t[\phi^{t+1}] \le \phi^t + \sum_r 2\eta_r^{t+1}p_r^{t+1}(-\phi_r^t + \tilde{\phi}_r^t) \\ +(\eta_r^{t+1})^2[\phi_r^t + E_t[\sum_{x\in A_r^{t+1}}\|x-\hat{c}_r^{t+1}\|^2]\end{aligned}$$

Now a key observation is that $p_r^{t+1} = 0$ if and only if cluster A_r^{t+1} is empty, i.e., degenerate. Since the degenerate clusters do not contribute to the k-means cost, we have $\sum_{r;p_r^{t+1}>0}\phi_r^t = \phi^t$, and similarly, $\sum_{r;p_r^{t+1}>0}\tilde{\phi}_r^t = \tilde{\phi}^t$. Therefore,

$$\begin{aligned}E_t[\phi^{t+1}] \le \phi^t - 2\min_{r,t;p_r^{t+1}>0}\eta_r^{t+1}p_r^{t+1}(\phi^t - \tilde{\phi}^t) \\ +(\eta_{\max}^{t+1})^2(E_t\sum_r\sum_{x\in A_r^{t+1}}\|x-\hat{c}_r^{t+1}\|^2 + \phi^t) \\ = \phi^t - 2\min_{r,t;p_r^{t+1}>0}\eta_r^{t+1}p_r^{t+1}(\phi^t-\tilde{\phi}^t) + (\eta_{\max}^{t+1})^2 6\phi^t\end{aligned}$$

where the last inequality is by Lemma 23. ☐

Lemma 15. *Suppose Assumption (A) holds. If we run Algorithm 1 on X with $\eta^t = \frac{c'}{t_o+t}$, and $t_o > 1$, with any initial set of k centroids $C^0 \in conv(X)$. Then for any $\delta > 0$, $\exists t$ s.t. $\Delta(C^t, C^*) \le \delta$ with $C^* := m(A^*)$ for some $A^* \in \{A^*\}_{[k]}$.*

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

Case 1: $m(A^{t+1}) \in \{C^*\}_{k^t}$

Then

$$\Delta(C^t, m(A^{t+1})) > \delta$$

by our assumption.

Case 2: $m(A^{t+1}) \notin \{C^*\}_{k^t}$

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

$$\Delta(C^t, m(A^{t+1})) \ge r_{\min}\phi(m(A^{t+1}))$$

So for both cases,

$$\Delta(C^t, m(A^{t+1})) \ge \min\{\delta, r_{\min}\phi_{opt}\}$$

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

$$\begin{aligned}E[\phi^{t+1} - \phi^t|F_t] \\ \le -\frac{2c'\min_{r\in[k];p_r^{t+1}(m)>0}p_r^{t+1}(m)}{t+1+t_o}\phi^t(1-\frac{\tilde{\phi}^t}{\phi^t}) \\ +(\frac{c'}{t+1+t_o})^2 6\phi_{\max}\end{aligned}$$

Note for $p_r^{t+1}(m) > 0$, by the discrete nature of the dataset, $\frac{n_r^{t+1}}{n} \ge \frac{1}{n}$, therefore,

$$\min_{r\in[k];p_r^t(m)>0} p_r^t(m) \ge 1 - (1-\frac{1}{n})^m \ge 1 - e^{-\frac{m}{n}}$$

Also note

$$\begin{aligned}\phi^t - \tilde{\phi}^t = \sum_{r\in[k']}\sum_{x\in A_r^{t+1}}\|x-C^t\|^2 - \|x-m(A_r^{t+1})\|^2 \\ = \sum_r \|c_r^t - m(A_r^{t+1})\|^2 n_r^{t+1} = \Delta(C^t, m(A^{t+1})) \ge \delta_o\end{aligned}$$

Then $\forall t \ge 1$,

$$\begin{aligned}E[\phi^{t+1}] - E[\phi^t] \\ \le -\frac{2c'(1-e^{-\frac{m}{n}})}{t+1+t_o}\delta_o + \frac{6\phi_{\max}(c')^2}{(t+1+t_o)^2}\end{aligned}$$

Summing up all inequalities,

$$\begin{aligned}E[\phi^{t+1}] - E[\phi^0] \\ \le -2c'(1-e^{-\frac{m}{n}})\delta_o \ln\frac{t+t_o+1}{t_o} + \frac{6\phi_{\max}(c')^2}{t_o-1}\end{aligned}$$

Since t is unbounded and $\ln \frac{t+t_o+1}{t_o}$ increases with t while $\frac{6\phi_{\max}(c')^2}{t_o-1}$ 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, $A^* \in \{A^*\}_{[k]}$. We call this event G; formally,

$$G := \{\exists T \geq 1, \exists A^* \in \{A^*\}_{[k]}, \text{ s.t. } A^t = A^*, \forall t \geq T\}$$

Second, we want to establish the $O(\frac{1}{t})$ 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 τ:

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

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

$$Pr(\{\tau < \infty\}) = Pr(\{\tau \in \mathbb{N}\}) = Pr(\cup_{T\geq 0}\{\tau = T\}) = 1 \quad (13)$$

Fixing τ, we denote the stationary clustering that the algorithm "hits" by

$$A^*(\tau) := \arg\min_{A^* \in \{A^*\}_{[k]}} \Delta(C^\tau, m(A^*))$$

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

We will prove a subset $G_o \subset G$ holds with high probability. To do this, we construct G_o as a union of disjoint events determined by the realization of τ and $A^*(\tau)$: we define events

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

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

$$G_o := \cup_{A^* \in \{A^*\}_{[k]}} G(A^*)$$

$G_o \subset G$ since the event $\Delta^t \leq r_{\min}\phi^*$ implies $A^t = A^*$.

Proof of Theorem 1. Fix any (T, A^*), conditioning on $\{\tau = T\} \cap \{A^*(\tau) = A^*\}$, since we have

$$c' > \frac{\phi_{\max}}{(1 - e^{-\frac{m}{n}}) r_{\min}\phi_{opt}}$$

We can envoke Lemma 16 to get $\forall t < T$,

$$E\{\phi^t - \phi(A^*) | G_T(A^*)\} = O(\frac{1}{t}) \quad (14)$$

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 $b_o = r_{\min}$, $\alpha = 0$, $p^*_{\min} \geq \frac{1}{n}$. Thus, for any

$$m \geq 1$$

$$c' > \frac{\beta}{2(1 - e^{\frac{4m}{5n}})} \quad \text{with} \quad \beta \geq 2$$

and

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

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

$$\begin{aligned} &Pr(\{\forall t \geq T, \Delta^t \leq r_{\min}\phi^*\} | \{\tau = T\} \cap \{A^*(\tau) = A^*\}) \\ &\quad = P(\Omega_\infty | \{\tau = T\} \cap \{A^*(\tau) = A^*\}) \geq 1 - \delta \quad (15) \end{aligned}$$

and by the second statement of Theorem 3, $\forall t > T$,

$$\begin{aligned} &E\{\phi^t - \phi(A^*) | \Omega_t, \{\tau = T\} \cap \{A^*(\tau) = A^*\}\} \\ &\leq E\{\Delta(C^t, C^*) | \Omega_t, \{\tau = T\} \cap \{A^*(\tau) = A^*\}\} = O(\frac{1}{t}) \end{aligned}$$

where the first inequality is by Lemma 18. Since $\Omega_\infty \subset \Omega_t$, $\forall t \geq 0$, this implies

$$\begin{aligned} &E\{\Delta(C^t, C^*) | \Omega_\infty, \{\tau = T\} \cap \{A^*(\tau) = A^*\}\} \\ &\quad = E\{\Delta(C^t, C^*) | G_T(A^*)\} = O(\frac{1}{t}) \quad (16) \end{aligned}$$

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

$$\begin{aligned} Pr\{G\} \geq Pr\{G_o\} &= Pr\{\cup_{T\geq 0} \cup_{A^* \in \{A^*\}_{[k]}} G_T(A^*)\} \\ &= \sum_{T\geq 0, A^* \in \{A^*\}_{[k]}} Pr\{G_T(A^*)\} \end{aligned}$$

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

$$\begin{aligned} &\sum_{T\geq 0, A^* \in \{A^*\}_{[k]}} Pr\{G_T(A^*)\} \\ &= \sum_{T, A^*} Pr\{\Omega_\infty | \{\tau = T\} \cap \{A^*(\tau) = A^*\}\} \\ &\qquad Pr(\{\tau = T\} \cap \{A^*(\tau) = A^*\}) \\ &\geq (1 - \delta) \sum_{T, A^*} Pr(\{\tau = T\} \cap \{A^*(\tau) = A^*\}) \\ &= (1 - \delta) Pr\{\cup_T \cup_{A^*} \{\tau = T\} \cap \{A^*(\tau) = A^*\}\} \\ &= (1 - \delta) Pr\{\cup_{T\geq 0}\{\tau = T\}\} = 1 - \delta \end{aligned}$$

where the inequality is by (15), and the last two equalities are due to the finiteness of $\{A^*\}_{[k]}$ and by (13), respectively. Therefore, $Pr\{G\} \geq 1 - \delta$, which completes the proof of the first statement. In addition,

$$\begin{aligned} &Pr\{\cup_{A^* \in \{A^*\}_{[k]}} G(A^*)\} \\ &= Pr\{\cup_{T\geq 0, A^* \in \{A^*\}_{[k]}} \Omega_\infty \cap \{\tau = T\} \cap \{A^*(\tau) = A^*\}\} \\ &\geq 1 - \delta \end{aligned}$$

which proves the second statement. Finally, combining inequalities (14) and (16), we have $\forall \geq 1$ and $\forall t \geq 1$,

$$E\{\phi^t - \phi(A^*)|G_T(A^*)\} = O(\frac{1}{t})$$

Since the quantity $\phi^t - \phi(A^{**})$ is independent of T, we reach the conclusion

$$E\{\phi^t - \phi(A^*)|G(A^*)\} = O(\frac{1}{t})$$

$\square$

Lemma 16. *Suppose the assumptions and settings in Theorem 1 hold, conditioning on any $G_T(A^*)$, we have $\forall 1 \leq t < T$,*

$$E\{\phi^t - \phi(A^*)|G_T(A^*)\} = O(\frac{1}{t})$$

Proof. First observe that conditioning on the event $G_T(A^*)$, $\Delta(C^t, C^*) > \frac{1}{2} r_{\min}\phi^*$, $\forall t < T$. 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

$$\phi^t - \tilde{\phi}^t > \min\{\frac{1}{2}r_{\min}, r_{\min}\}\tilde{\phi}^t = \frac{1}{2}r_{\min}\tilde{\phi}^t$$

Proceeding as in Lemma 15, we have conditioning on $G_T(A^*)$,

$$E[\phi^t|G_T(A^*)]$$
$$\leq E[\phi^{t-1}|G_T(A^*)]\{1 - \frac{2c' \min_{r\in[k];p_r^t(m)>0} p_r^t(m)}{t+t_o}\frac{r_{\min}\phi_{opt}}{2\phi_{\max}}\} + (\frac{c'}{t+t_o})^2 6\phi_{\max}$$

since $\forall t \geq 1$,

$$1 - \frac{\tilde{\phi}^t}{\phi^t} \geq \frac{r_{\min}}{2}\frac{\tilde{\phi}^t}{\phi^t} \geq \frac{r_{\min}}{2}\frac{\phi_{opt}}{\phi_{\max}}$$

Now, since we set

$$c' > \frac{\phi_{\max}}{(1-e^{-\frac{m}{n}})r_{\min}\phi_{opt}}$$

we have

$$\begin{aligned}
&2c' \min_{r\in[k];p_r^t(m)>0} p_r^t(m)\frac{r_{\min}\phi_{opt}}{2\phi_{\max}} \\
&\geq 2c'(1-(1-\frac{1}{n})^m)\frac{r_{\min}\phi_{opt}}{2\phi_{\max}} \\
&\geq 2c'(1-e^{-\frac{m}{n}})\frac{r_{\min}\phi_{opt}}{2\phi_{\max}} \\
&> 2\frac{\phi_{\max}}{(1-e^{-\frac{m}{n}})r_{\min}\phi_{opt}}(1-e^{-\frac{m}{n}})\frac{r_{\min}\phi_{opt}}{2\phi_{\max}} > 1
\end{aligned}$$

Applying Lemma 26 with

$$a := 2c' \min_{r\in[k];p_r^t(m)>0} p_r^t(m)\frac{r_{\min}\phi_{opt}}{2\phi_{\max}} > 1$$

$$b := \frac{6(c')^2\phi_{\max}}{(t_o+t)^2}$$

We conclude that $\forall 1 \leq t < T$,

$$E[\phi^t|G_T(A^*)] \leq \frac{t_o+1}{t_o+t+1}\phi^o + \frac{b}{a-1}(\frac{t_o+2}{t_o+1})^{a+1}\frac{1}{t_o+t+1}$$

Subtracting $\phi(A^*)$ from both sides of the equation, we get

$$\begin{aligned}
E[\phi^t - \phi(A^*)|G_T(A^*)] &\leq \frac{t_o+1}{t_o+t+1}(\phi^o - \phi(A^*)) \\
&+ \frac{b}{a-1}(\frac{t_o+2}{t_o+1})^{a+1}\frac{1}{t_o+t+1} = O(\frac{1}{t})
\end{aligned}$$

$\square$

Proofs leading to Theorem 2

Here, we additionally define two quantities that characterizes C^*: Let $A^* = v(C^*)$, we use $p^*_{\min} := \min_{r\in[k]} \frac{n_r^*}{n}$ to characterize the fraction of the smallest cluster in A_* to the entire dataset. We use $w_r := \frac{\frac{\phi_*^r}{n_r^*}}{\max_{x\in A_r^*}\|x-c_r^*\|^2}$ to characterize the ratio between average and maximal "spread" of cluster A_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 $\|c_r^* - c_s^*\|$, $\forall r, s \neq r$. Let $x \in A_r^* \cap A_s^t$. Split x into its projection on the line joining c_r^* and c_s^*, and its orthogonal component:

$$x = \frac{1}{2}(c_r^* + c_s^*) + \lambda(c_r^* - c_s^*) + u \tag{17}$$

with $u \perp c_r^* - c_s^*$. Note λ measures the ratio between departure of the projected point from the mid-point of c_r^* and c_s^* and the norm $\|c_r^* - c_s^*\|$. By minimality of our definition of margin Δ_{rs},

$$\|\bar{x} - \frac{1}{2}(c_r^* + c_s^*)\| = \lambda\|c_r^* - c_s^*\| \geq \frac{1}{2}\Delta_{rs} \tag{18}$$

In addition, since c_r^* is the mean of A_r^*, we know there exists $x \in A_r^*$ such that $\bar{x}$ falls outside of the line segment $c_r^* - c_s^*$ (or exactly on c_r^* in the special case where all points projects on c_r^*). Similar holds for c_s^*. Thus,

$$\|c_r^* - c_s^*\| \geq \Delta_{rs} \geq f(\alpha)\sqrt{\phi^*}(\frac{1}{\sqrt{n_r^*}} + \frac{1}{\sqrt{n_s^*}}) \tag{19}$$

Lemma 17 (Theorem 5.4 of [12])**.** *Suppose (X, C^*) satisfies (B). If $\forall r \in [k], s \neq r$, $\Delta_r^t + \Delta_s^t \leq \frac{\Delta_{rs}}{16}$. Then for any $s \neq r$, $|A_r^* \cap A_s^t| \leq \frac{b^2}{f(\alpha)}$, where $b \geq \max_{r,s}\frac{\Delta_r^t + \Delta_s^t}{\Delta_{rs}}$.*

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_r^t, c_s^t is closer to s, we have

$$x(c_s^t - c_r^t) \geq \frac{1}{2}(c_s^t - c_r^t)(c_s^t + c_r^t)$$

Substituting (17) into the inequality above,

$$\frac{1}{2}(c_r^* + c_s^*)(c_s^t - c_r^t) + \lambda(c_r^* - c_s^*)(c_s^t - c_r^t) + u(c_s^t - c_r^t) \geq \frac{1}{2}(c_s^t - c_r^t)(c_s^t + c_r^t) \tag{20}$$

Since $u \perp c_r^* - c_s^*$, let $\Delta = \Delta_s^t + \Delta_r^t$. We have

$$u(c_s^t - c_r^t) = u(c_s^t - c_s^* - (c_r^t - c_r^*)) \leq \|u\|\Delta$$

Rearranging (20), we have

$$\frac{1}{2}(c_r^* + c_s^* - c_s^t - c_r^t)(c_s^t - c_r^t) + \lambda(c_r^* - c_s^*)(c_s^t - c_r^t) + u(c_s^t - c_r^t) \geq 0$$
$$\equiv \frac{\Delta^2}{2} + \frac{\Delta}{2}\|c_r^* - c_s^*\| - \lambda\|c_r^* - c_s^*\|^2 + \lambda\Delta\|c_r^* - c_s^*\| + \|u\|\Delta \geq 0$$

Therefore,

$$\|x - c_r^*\| = \|(\frac{1}{2} - \lambda)(c_s^* - c_r^*) + u\| \geq \|u\|$$
$$\geq \frac{\lambda}{\Delta}\|c_r^* - c_s^*\|^2 - \frac{\Delta}{2} - \frac{1}{2}\|c_r^* - c_s^*\| - \lambda\|c_r^* - c_s^*\| \geq \frac{\Delta_{rs}\|c_r^* - c_s^*\|}{64\Delta}$$

where the last inequality is by our assumption that $\Delta \leq \frac{\Delta_{rs}}{16}$, and $\lambda \geq \frac{\Delta_{rs}}{2\|c_r^* - c_s^*\|}$ by (18). By previous inequality and our assumption on f, [3] for all $s \neq r$

$$|A_r^* \cap A_s^t| \frac{\Delta_{rs}^2 \|c_r^* - c_s^*\|^2}{f\Delta^2} \leq \sum_{x \in A_r^* \cap A_s^t} \|x - c_r^*\|^2$$

So $|A_r^* \cap A_s^t| \leq \sum_{x \in A_r^* \cap A_s^t} \|x - c_r^*\|^2 \frac{f(\Delta_r^t + \Delta_s^t)^2}{\Delta_{rs}^2 \|c_r^* - c_s^*\|^2} \leq \frac{fb^2}{f^2\phi^*(\frac{1}{n_r^*})}(\sum_{A_r^* \cap A_s^t} \|x - c_r^*\|^2)$, where the second inequality is by (19). That is, $\frac{|A_r^* \cap A_s^t|}{n_r^*} \leq \frac{b^2}{f\phi^*}\sum_{A_r^* \cap A_s^t}\|x - c_r^*\|^2$. Similarly, for all $s \neq r$, $\frac{|A_s^* \cap A_r^t|}{n_r^*} \leq \frac{b^2}{f\phi^*}\sum_{A_s^* \cap A_r^t}\|x - c_s^*\|^2$ Summing over all $s \neq r$, $\frac{|A_r \triangle A_r^*|}{n_r^*} = \rho_{out} + \rho_{in} \leq \frac{b^2}{f\phi^*}\phi^* = \frac{b^2}{f}$. ◻

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*

$$\phi(C) - \phi^* \leq \min_\pi \sum_r n_r^* \|c_{\pi(r)} - c_r^*\|^2 = \Delta(C, C^*)$$

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 π,

$$\phi(C) - \phi^* = \sum_s \sum_{x \in A_s} \|x - c_s\|^2 - \sum_r \sum_{x \in A_r^*} \|x - c_r^*\|^2$$
$$\leq \sum_r \sum_{x \in A_r^*} \|x - c_{\pi(r)}\|^2 - \sum_r \sum_{x \in A_r^*} \|x - c_r^*\|^2$$
$$= \sum_r n_r^* \|c_{\pi(r)} - c_r^*\|^2$$

where the last inequality is by optimality of clustering assignment 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 π, it must holds for $\min_\pi \sum_r n_r^* \|c_{\pi(r)} - c_r^*\|^2$, which completes the proof. ◻

Proofs regarding seeding guarantee

Lemma 19 (Theorem 4 of [19])**.** *Suppose (X, C^*) satisfies (B). If we obtain seeds from Algorithm 2, then*

$$\Delta(C^0, C^*) \leq \frac{1}{2}\frac{f(\alpha)^2}{16^2}\phi^*$$

with probability at least $1 - m_o \exp(-2(\frac{f(\alpha)}{4} - 1)^2 w_{\min}^2) - k\exp(-m_o p_{\min}^*)$.

Proof. First recall that, as in (19), assumption (B) implies center-separability assumption in Definition 1 of [19], i.e.

$$\forall r \in [k], s \neq r, \|c_r^* - c_s^*\| \geq f(\alpha)\sqrt{\phi^*}(\frac{1}{\sqrt{n_r^*}} + \frac{1}{\sqrt{n_s^*}})$$

with $f(\alpha) \geq \max_{r \in [k], s \neq r} \frac{n_r^*}{n_s^*}$. [4] Applying Theorem 4 of [19] with $\mu_r = c_r^*$ and $\nu_r = c_r^0$, we get $\forall r \in [k]$, $\|c_r^0 - c_r^*\| \leq \frac{\sqrt{f(\alpha)}}{2}\sqrt{\frac{\phi_r^*}{n_r^*}}$ with probability at least $1 - m_o\exp(-2(\frac{f(\alpha)}{4} - 1)^2 w_{\min}^2) - k\exp(-m_o p_{\min}^*)$. Summing over all r, the previous event implies $\sum_r n_r^* \|c_r^0 - c_r^*\|^2 \leq \frac{f(\alpha)}{4}\phi^* \leq \frac{1}{2}\frac{f(\alpha)^2}{16^2}\phi^*$, where the last inequality is by the assumption that $f \geq 64^2$ in (B). ◻

Lemma 20. *Assume the conditions Lemma 19 hold. For any $\xi > 0$, if in addition,*

$$f(\alpha) \geq 5\sqrt{\frac{1}{2w_{\min}}\ln(\frac{2}{\xi p_{\min}^*}\ln\frac{2k}{\xi})}$$

If we obtain seeds from Algorithm 2 choosing

$$\frac{\ln\frac{2k}{\xi}}{p_{\min}^*} < m_o < \frac{\xi}{2}\exp\{2(\frac{f(\alpha)}{4} - 1)^2 w_{\min}^2\}$$

Then $\Delta(C^0, C^*) \leq \frac{1}{2}\frac{f(\alpha)^2}{16^2}\phi^*$ with probability at least $1 - \xi$.

Proof. By Lemma 19, a sufficient condition for the success probability to be at least $1 - \xi$ is:

$$m_o \exp(-2(\frac{f(\alpha)}{4} - 1)^2 w_{\min}^2) \leq \frac{\xi}{2}$$

and

$$k\exp(-m_o p_{\min}^*) \leq \frac{\xi}{2}$$

This translates to requiring

$$\frac{1}{p_{\min}^*}\ln\frac{2k}{\xi} \leq m_o \leq \frac{\xi}{2}\exp(2(\frac{f(\alpha)}{4} - 1)^2 w_{\min}^2)$$

[3] We use f as a shorthand for $f(\alpha)$ in the subsequent proof.

[4] note: "α" in [19] is defined as $\min_{r \in [k], s \neq r} \frac{n_r^*}{n_s^*}$, which is not to be confused with our "α".

Note for this inequality to be possible, we also need $\frac{1}{p^*_{\min}} \ln \frac{2k}{\xi} \le \frac{\xi}{2} \exp(2(\frac{f(\alpha)}{4} - 1)^2 w^2_{\min})$, imposing a constraint on $f(\alpha)$. Taking logarithm on both sides and rearrange, we get

$$(\frac{f(\alpha)}{4} - 1)^2 \ge \frac{1}{2w_{\min}} \ln(\frac{2}{\xi p^*_{\min}} \ln \frac{2k}{\xi})$$

This satisfied since $f(\alpha) \ge 5\sqrt{\frac{1}{2w_{\min}} \ln(\frac{2}{\xi p^*_{\min}} \ln \frac{2k}{\xi})}$. □

Proof of Theorem 2. By Proposition 1, (X, C^*) satisfying (B) implies C^* is $(\frac{f(\alpha)^2}{16^2}, \alpha)$-stable. Let $b_0 := \frac{f(\alpha)^2}{16^2}$, and we denote event $F := \{\Delta(C^0, C^{opt}) \le \frac{1}{2} b_0 \phi^*\}$. Since $f(\alpha) \ge 5\sqrt{\frac{1}{2w_{\min}} \ln(\frac{2}{\xi p^*_{\min}} \ln \frac{2k}{\xi})}$, and $\frac{\log \frac{2k}{\xi}}{p^*_{\min}} < m_o < \frac{\xi}{2} \exp\{2(\frac{f(\alpha)}{4} - 1)^2 w^2_{\min}\}$, we can apply Lemma 20 to get

$$Pr\{F\} \ge 1 - \xi$$

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

$$E\{\Delta^t | \Omega_t, F\} = O(\frac{1}{t}) \text{ and } Pr\{\Omega_t | F\} \ge 1 - \delta$$

So

$$Pr\{\Omega_t \cap F\} = Pr\{\Omega_t | F\} Pr\{F\} \ge (1 - \delta)(1 - \xi)$$

Finally, using Lemma 18, and letting $G_t := \Omega_t \cap F$, 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 η^t_r is equivalent to online k-means [6] and mini-batch k-means [18].

Claim 1. *In Algorithm 1, if we set a counter for* $\hat{N}^t_r := \sum_{i=1}^t \hat{n}^i_r$ *and if we set the learning rate* $\eta^t_r := \frac{\hat{n}^t_r}{\hat{N}^t_r}$, *then provided the same random sampling scheme is used,*

1. *When mini-batch size $m = 1$, the update of Algorithm 1 is equivalent to that described in [Section 3.3, [6]].*
2. *When $m > 1$, the update of Algorithm 1 is equivalent to that described from line 3 to line 14 in [Algorithm 1, [18]] with mini-batch size m.*

Proof. For the first claim, we first re-define the variables used in [Section 3.3, [6]]. We substitute index k in [6] with r used in Algorithm 1. For any iteration t, we define the equivalence of definitions: $s \leftarrow x_i$, $c^t_r \leftarrow w_k$, $\hat{n}^t_r \leftarrow \Delta n_k$, $\hat{N}^t_r \leftarrow n_k$. According to the update rule in [6], $\Delta n_k = 1$ 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 [6] 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

$$c^t_r = c^{t-1}_r + \frac{1}{\hat{N}^t_r}(s - c^{t-1}_r) 1_{\{\hat{n}^t_r = 1\}}$$

Now the update defined by Algorithm 1 with $m = 1$ and $\eta^t_r = \frac{\hat{n}^t_r}{\hat{N}^t_r}$ is:

$$\begin{aligned} c^t_r &= c^{t-1}_r + \eta_r^t(\hat{c}^t_r - c^{t-1}_r) 1_{\{\hat{n}^t_r \neq 0\}} \\ &= c^{t-1}_r + \frac{\hat{n}^t_r}{\hat{N}^t_r}(s - c^{t-1}_r) 1_{\{\hat{n}^t_r = 1\}} \\ &= c^{t-1}_r + \frac{1}{\hat{N}^t_r}(s - c^{t-1}_r) 1_{\{\hat{n}^t_r = 1\}} \end{aligned}$$

since $\hat{n}^t_r$ 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, [18]]. We substitute their index of time i with t in Algorithm 1. We define the equivalence of definitions: $m \leftarrow b$, $S^t \leftarrow M$, $s \leftarrow x$, $c^{t-1}_{I(s)} \leftarrow d[x]$, $c^{t-1}_r \leftarrow c$.

At iteration t, we let $v[c^{t-1}_r]_t$ denote the value of counter $v[c]$ upon completion of the loop from line 9 to line 14 for each center c, then $\hat{N}^t_r \leftarrow v[c^{t-1}_r]_t$. Since according to Lemma 22, from line 9 to line 14, the updated centroid c^t_r after iteration t is

$$c^t_r = \frac{1}{v[c^{t-1}_r]_t} \sum_{s \in \cup_{i=1}^t S^i_r} s = \frac{1}{\hat{N}^t_r} \sum_{s \in \cup_{i=1}^t S^i_r} s$$

This implies

$$\begin{aligned} c^t_r - c^{t-1}_r &= \frac{1}{\hat{N}^t_r} \sum_{s \in \cup_{i=1}^t S^i_r} s - c^{t-1}_r \\ &= \frac{1}{\hat{N}^t_r}[\sum_{s \in S^t_r} s + \sum_{s' \in \cup_{i=1}^{t-1} S^i_r} s'] - c^{t-1}_r \\ &= \frac{1}{\hat{N}^t_r}[\sum_{s \in S^t_r} s + \hat{N}^{t-1}_r c^{t-1}_r] - c^{t-1}_r \\ &= -\frac{\hat{n}^t_r}{\hat{N}^t_r} c^{t-1}_r + \frac{\hat{n}^t_r}{\hat{N}^t_r} \frac{\sum_{s \in S^t_r} s}{\hat{n}^t_r} = -\eta^t_r c^{t-1}_r + \eta^t_r \hat{c}^t_r \end{aligned}$$

Hence, the updates in Algorithm 1 and line 4 to line 14 in [Algorithm 1, [18]] are equivalent. □

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

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

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 repeatdly 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 $T=1$, $w_1 = (1-1)w_0 + g_1 = \frac{1}{1}\sum_{t=1}^{1} g_t$. So the claim holds for $T=1$.

Suppose the claim holds for T, then for $T+1$, by the update rule

$$\begin{aligned} w_{T+1} &= (1-\frac{1}{T+1})w_T + \frac{1}{T+1}g_{T+1} \\ &= (1-\frac{1}{T+1})\frac{1}{T}\sum_{t=1}^{T} g_t + \frac{1}{T+1}g_{T+1} \\ &= \frac{T}{T+1}\frac{1}{T}\sum_{t=1}^{T} g_t + \frac{1}{T+1}g_{T+1} \\ &= \frac{1}{T+1}\sum_{t=1}^{T+1} g_t \end{aligned}$$

So the claim holds for any $T \geq 1$. □

Lemma 23. *$\forall t \geq 1$, conditioning on F_t, the noise term (10) is upper bounded by $B_1 := 5\phi^t$.*

Proof. Since

$$\|x - \hat{c}_r^{t+1}\|^2 \leq 2\|x - c_r^t\|^2 + 2\|c_r^t - \hat{c}_r^{t+1}\|^2$$

We have

$$\begin{aligned} &E[\sum_r \sum_{x \in A_r^{t+1}} \|x - \hat{c}_r^{t+1}\|^2 + \phi^t | F_t] \\ &\leq 2\sum_r \sum_{x \in A_r^{t+1}} \|x - c_r^t\|^2 \\ &+2\sum_r \sum_{x \in A_r^{t+1}} E[\|c_r^t - \hat{c}_r^{t+1}\|^2 | F_t] + \phi^t \end{aligned}$$

Now,

$$E[\|c_r^t - \hat{c}_r^{t+1}\|^2 | F_t] \leq E\frac{\sum_{s \in S_r^t} \|c_r^t - s\|^2}{|S_r^t|} = \frac{\phi_r^t}{n_r^t}$$

where S_r^t is the sampled from A_r^t in Algorithm 1, and the inequality is by convexity of l_2-norm. Substituting this into the previous inequality completes the proof. □

Lemma 24. *Suppose C^* is (b_o, α)-stable. Conditioning on Ω_i, we have, The terms (11), and (12), for $t = i$, are upper bounded by $B := 4(b_o+1)n\phi^*$.*

Proof. Conditioning on Ω_i,

$$\Delta^{i-1} \leq b_o\phi^*$$

By Lemma 18, we also have

$$\phi^{i-1} - \phi^* \leq \Delta^{i-1} \leq b_o\phi^*$$

By Cauchy-Schwarz,

$$\begin{aligned} &\sum_r n_r^* \langle c_r^{i-1} - c_r^*, \hat{c}_r^i - c_r^{i-1}\rangle \\ &\leq \sqrt{\sum_r n_r^* \|c_r^{i-1} - c_r^*\|^2}\sqrt{\sum_r n_r^* \|\hat{c}_r^i - c_r^{i-1}\|^2} \end{aligned}$$

Now, since $\hat{c}_r^i$ is the mean of a subset of A_r^i,

$$\|\hat{c}_r^i - c_r^{i-1}\|^2 \leq \phi_r^{i-1}$$

Hence

$$\sum_r n_r^* \|\hat{c}_r^i - c_r^{i-1}\|^2 \leq n\phi^{i-1}$$

On the other hand,

$$\begin{aligned} \sum_r n_r^* \|c_r^{i-1} - E[\hat{c}_r^i | F_{i-1}]\|^2 &= \sum_r n_r^* \|c_r^{i-1} - m(A_r^i)\|^2 \\ &\leq n\sum_r \phi(c_r^{i-1}) - \phi(m(A_r^i)) \\ &= n[\phi^{i-1} - \phi(m(A^i))] \leq n(\phi^{i-1} - \phi^*) \end{aligned}$$

Now we first bound (11):

$$\begin{aligned} &\sum_r n_r^* \langle c_r^{i-1} - c_r^*, \hat{c}_r^i - E[\hat{c}_r^i | F_{i-1}]\rangle \\ &= \sum_r n_r^* \langle c_r^{i-1} - c_r^*, \hat{c}_r^i - c_r^{i-1}\rangle \\ &+ \sum_r n_r^* \langle c_r^{i-1} - c_r^*, c_r^{i-1} - E[\hat{c}_r^i | F_{i-1}]\rangle \\ &\leq \sqrt{\Delta^{i-1}}\sqrt{n\phi^{i-1}} + \sqrt{\Delta^{i-1}}\sqrt{n(\phi^{i-1} - \phi^*)} \\ &\leq \sqrt{b_o\phi^*}\sqrt{n(b_o+1)\phi^*} + \sqrt{n}b_o\phi^* \leq 2(b_o+1)\sqrt{n}\phi^* \end{aligned}$$

To bound (12),

$$\begin{aligned} &\sum_r n_r^* \|\hat{c}_r^i - c_r^*\|^2 \\ &\leq 2\sum_r n_r^* \|\hat{c}_r^i - c_r^{i-1}\|^2 + 2\sum_r n_r^* \|c_r^{i-1} - c_r^*\|^2 \\ &\leq 2n\phi^{i-1} + 2\Delta^{i-1} \leq 2n(b_o+1)\phi^* + 2b_o\phi^* \leq 4n(b_o+1)\phi^* \end{aligned}$$

□

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

Proof of Claim. By the update rule in Algorithm 1, c_r^{t+1} is a convex combination of c_r^t and $\hat{c}_r^{t+1}$, where $\hat{c}_r^{t+1}$ is the mean of a subset of X, and hence $\hat{c}_r^{t+1} \in conv(X)$. Since both c_r^t and $\hat{c}_r^{t+1}$ are in $conv(X)$, $c_r^{t+1} \in conv(X)$. □

Lemma 25 (technical lemma)**.** *For any fixed $b \in (1,2]$. If $C \geq \frac{b-1}{3}, \delta \leq \frac{1}{e}$, and $t \geq (\frac{3C}{b-1}\ln\frac{1}{\delta})^{\frac{2}{b-1}}$, then $t^{b-1} - 2C\ln t - C\ln\frac{1}{\delta} > 0$.*

Proof. Let $f(t) := t^{b-1} - 2C\ln t - C\ln\frac{1}{\delta}$. Taking derivative, we get $f'(t) = (b-1)t^{b-2} - \frac{2C}{t} \geq 0$ when $t \geq (\frac{2C}{b-1})^{\frac{1}{b-1}}$. Since $\ln\frac{1}{\delta}\frac{3C}{b-1} \geq \frac{3C}{b-1} \geq 1$, $(\ln\frac{1}{\delta}\frac{3C}{b-1})^{\frac{2}{b-1}} \geq (\frac{2C}{b-1})^{\frac{1}{b-1}}$, it suffices to show $f((\ln\frac{1}{\delta}\frac{3C}{b-1})^{\frac{2}{b-1}}) > 0$ for our statement to hold. $f((\ln\frac{1}{\delta}\frac{3C}{b-1})^{\frac{2}{b-1}}) = (\ln\frac{1}{\delta}\frac{3C}{b-1})^2 - 2C\ln\{(\ln\frac{1}{\delta}\frac{3C}{b-1})^{\frac{2}{b-1}}\} - C\ln\frac{1}{\delta} = (\ln\frac{1}{\delta})^2\frac{9C^2}{(b-1)^2} - \frac{4C}{b-1}\ln(\ln\frac{1}{\delta}\frac{3C}{b-1}) - C\ln\frac{1}{\delta} = \frac{4C}{b-1}[\frac{\frac{3}{2}C}{b-1}\ln\frac{1}{\delta} - \ln(\frac{3C}{b-1}\ln\frac{1}{\delta})] + C\ln\frac{1}{\delta}[\frac{3C}{(b-1)^2} - 1] > 0$, 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. □

Lemma 26 (Lemma D1 of [4]). *Consider a nonnegative sequence $(u_t : t \geq t_o)$, such that for some constants $a, b > 0$ and for all $t > t_o \geq 0$, $u_t \leq (1 - \frac{a}{t})u_{t-1} + \frac{b}{t^2}$. Then, if $a > 1$,*

$$u_t \leq (\frac{t_o + 1}{t + 1})^a u_{t_o} + \frac{b}{a-1}(1 + \frac{1}{t_o + 1})^{a+1} \frac{1}{t+1}$$

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, `mnist` and `covtype`, a simulated dataset `gauss`, and add stochastic k-means with a constant learning rate. Instead of a 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 $\frac{b}{t}$ (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 $\frac{b}{t+t_o}$, 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 4, when t_o is set to be 60 or higher, the actual convergence can be simulated by (a proxy to) our theoretical bound[5], $\frac{\phi^o - \phi_{\min}}{t+t_o}$. Note the practical requirement on t_o is much more optimistic than the lower bound in Theorem 1, i.e.,

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

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 $\frac{\phi^T}{\phi_{batch}}$. ϕ^T is the k-means cost of stochastic k-means after T iterations, with $T = 20 \times E$, where E is a particular epoch length. ϕ_{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 `mnist, covtype, gauss, 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!

[5]The difference between their intercepts at the y-axis is caused by a constant factor.

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 $\frac{b}{t+t_o}$ is observed for all datasets, stochastic k-means seems to favor certain datasets to others. For example, on `rcv1`, almost $\frac{b}{t}$ (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_o to be one of $\{10, 60, 600, 6000\}$, which ever gives the lowest k-means cost.

	`covtype`		
k	E=60,flat,BBS,const	E=600,flat,BBS,const	E=6k,flat,BBS,const
10	0.93,0.92,0.93	0.99,0.93,0.99	1.03,1.01,1.03
50	1.13,1.12,1.13	1.02,1.03,1.02	1.01,1.01,1.01
100	1.15,1.10,1.15	1.05,1.07,1.05	1.02,1.01,1.02
	`mnist`		
10	1.07,1.07,1.07	1.02,1.02,1.02	1.03,1.02,1.03
50	1.15,1.15,1.15	1.06,1.07,1.06	1.02,1.02,1.02
100	1.18,1.18,1.18	1.07,1.06,1.07	1.02,1.02,1.02

`rcv1`		
k	E=60,flat,BBS,const	E=600,flat,BBS,const
10	1.03,1.03,1.03	1.02,1.02,1.02
50	1.06,1.06,1.06	1.06,1.06,1.06
100	1.09,1.09,1.09	1.07,1.07,1.07
`gauss`		
k	E=60,flat,BBS,const	E=600,flat,BBS,const
	1.05,1.07,1.05	1.03,1.03,1.03
	1.16,1.14,1.16	1.07,1.05,1.07
	1.11,1.11,1.11	1.02,1.02,1.02

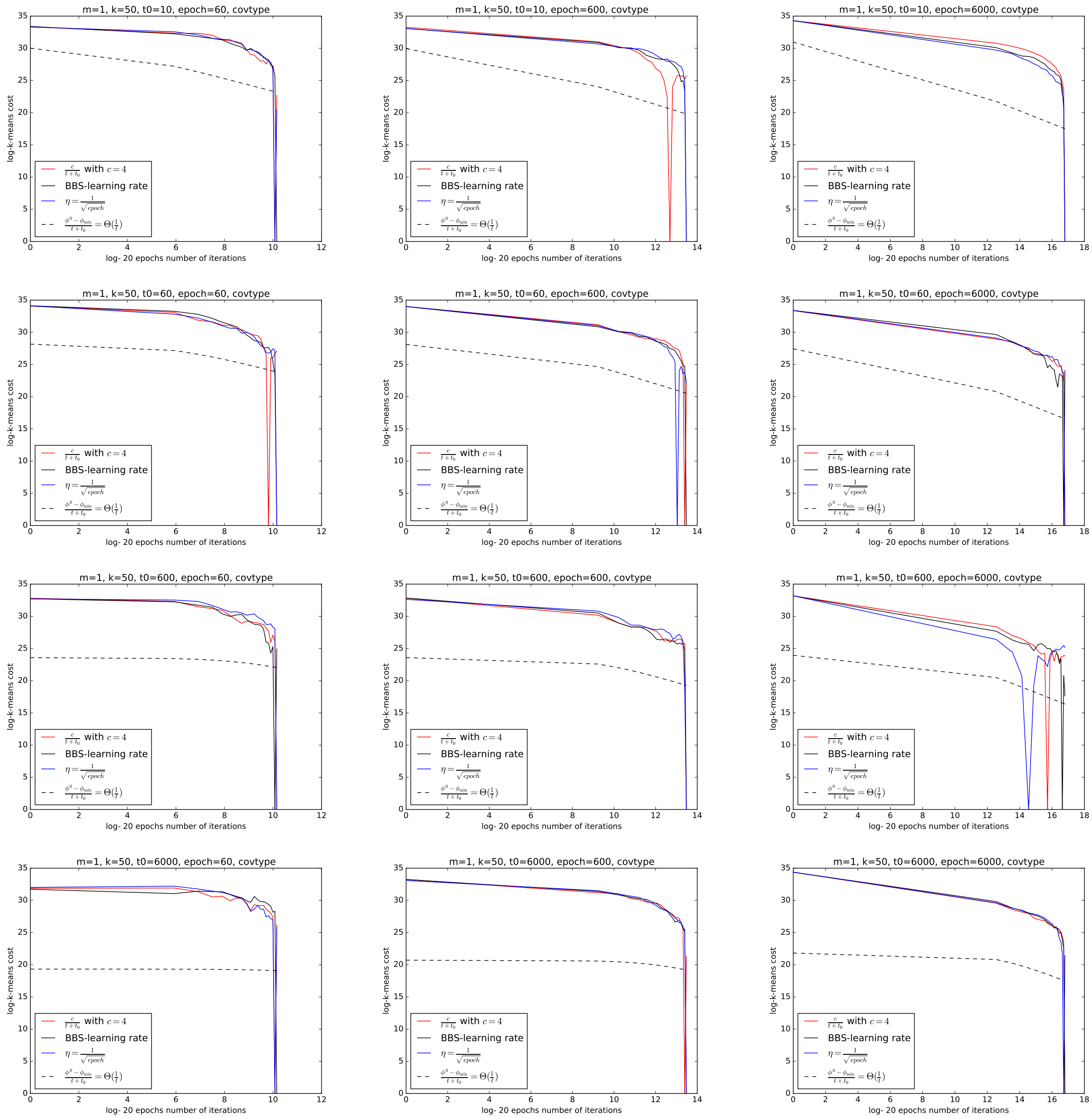

Figure 4: Experiments on `covtype`

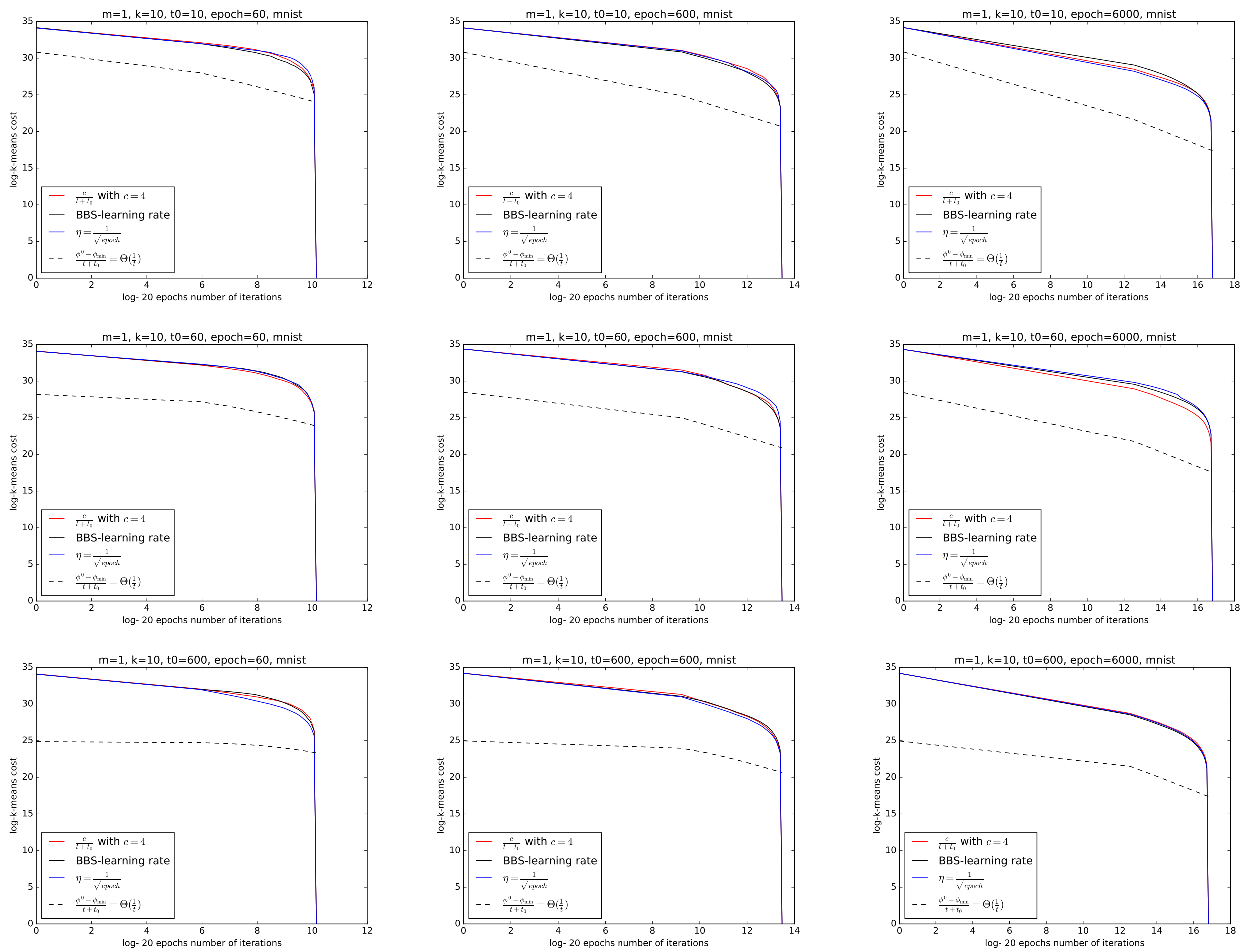

Figure 5: Experiments on `mnist`