Scalable constant k-means approximation via heuristics on well-clusterable data

Cheng Tang
tangch@gwu.edu
Department of Computer Science
George Washington University
Washington, DC, 22202

Claire Monteleoni
cmontel@gwu.edu
Department of Computer Science
George Washington University
Washington, DC, 22202

Supplementary of remaining proofs

Proof of Corollary 1

Proof. We first find a sufficient condition for Algorithm 1 to have a $1+\epsilon$-approximation. Note, as in the proof of Theorem 1, the approximation guarantee is upper bounded by $(\frac{1}{1-4\gamma})^2$, where $\gamma \leq \frac{\sqrt{f}}{2f}$. So to have a $1+\epsilon$-guarantee, it suffices to have $(\frac{1}{1-4\frac{\sqrt{f}}{2f}})^2 \leq 1+\epsilon$, which holds if $f = \Omega(\frac{1}{\epsilon^2})$. Now we find a sufficient condition for the success probability to be at least $1-\delta$. It suffices to require that $m\exp(-2(\frac{f}{4}-1)^2 w_{\min}^2) \leq \frac{\delta}{2}$ and $k\exp(-mp_{\min}) \leq \frac{\delta}{2}$. So we need $\frac{1}{p_{\min}}\log\frac{2k}{\delta} \leq m \leq \frac{\delta}{2}\exp(2(\frac{f}{4}-1)^2 w_{\min}^2)$. Note for this inequality to be possible, we also need $\frac{\delta}{2}\exp(2(\frac{f}{4}-1)^2 w_{\min}^2) \geq \frac{1}{p_{\min}}\log\frac{2k}{\delta}$, imposing an additional constraint on f. Taking $\log$ on both sides and rearrange, we get $(\frac{f}{4}-1)^2 \geq \frac{1}{2w_{\min}}\log(\frac{\frac{2}{\delta}\log\frac{2k}{\delta}}{p_{\min}})$. Thus, it is sufficient for a $1+\epsilon$-approximation to hold with probability at least $1-\delta$ if $f = \Omega\left(\sqrt{\log(\frac{\frac{1}{\delta}\log\frac{k}{\delta}}{p_{\min}})} + \frac{1}{\epsilon^2}\right)$, and we choose m to be in the interval $[\frac{1}{p_{\min}}\log\frac{2k}{\delta}, \frac{\delta}{2}\exp(2(\frac{f}{4}-1)^2 w_{\min}^2)]$. $\square$

Proof of Theorem 2

Proof. The proof mostly relies on our analysis of Lloyd's algorithm in [1]. First, Theorem 4 of [1], an analogous result to Theorem 3 here (the former holds w.r.t. $d_{rs}^*(f)$-center separability [1] instead of the weak center separability here), implies the upper bound on seeding $\|\mu_r - \nu_r^*\| \leq \frac{\sqrt{f}}{2}\sqrt{\frac{\phi_*^r}{n_r}} \leq \frac{\sqrt{f}}{2}\sqrt{\frac{\phi_*}{n_r}}, \forall r \in [k]$, satisfies the condition in Theorem 1[1]. Let $\{\nu_r^{fin}\}$ denote the set of k centroids obtained by running Lloyd's algorithm until convergence with seeding $\{\nu_r^*\}$ obtained from Algorithm 1. Applying Theorem 1 [1] repeatedly, we get $\max_r \|\nu_r^{fin} - \mu_r\| \leq \frac{128}{9f}\sqrt{\frac{\phi_*}{n_r}}$. Now we can proceed using the proof of Theorem 1 in this paper, only substituting γ with a tighter bound, that is, $\gamma \leq \frac{\frac{128}{9f}}{f} = O(\epsilon)$ when $f = \Omega(\frac{1}{\sqrt{\epsilon}})$, which guarantees $\frac{1}{(1-4\gamma)^2} \leq 1+\epsilon$. So the dependence of f on ϵ is now $\Omega(\frac{1}{\sqrt{\epsilon}})$. $\square$

Proof of Lemma 1

Proof. $\|x-\mu_s\| \geq \|x-\nu_s^*\| - \|\mu_s - \nu_s^*\| \geq \frac{1}{2}\|\nu_s^* - \nu_r^*\| - \|\mu_s - \nu_s^*\|$, since x is closer to ν_r^* by the Voronoi partition induced by $\{\nu_i^*, i \in [k]\}$. Now $\|\nu_r^* - \nu_s^*\| = \|\nu_r^* - \mu_r + \mu_r - \mu_s + \mu_s - \nu_s^*\| \geq \|\mu_r - \mu_s\| - \|\nu_s^* - \mu_s\| - \|\nu_r^* - \mu_r\| \geq (1-2\gamma)\|\mu_r - \mu_s\|$, by definition of γ. This implies $\|x-\mu_s\| \geq (\frac{1}{2}-\gamma)\|\mu_r - \mu_s\| - \|\mu_s - \nu_s^*\| \geq (\frac{1}{2}-2\gamma)\|\mu_r - \mu_s\|$ where $\|\mu_r - \mu_s\| \geq \frac{1}{\gamma}\|\nu_s^* - \mu_s\|$

and $\|\mu_r-\mu_s\| \geq \frac{1}{\gamma}\|\nu_r^*-\mu_r\|$. Finally, $\|x-\mu_r\| \leq \|\mu_r-\nu_r^*\|+\|x-\nu_r^*\| \leq \|\mu_r-\nu_r^*\|+\|x-\nu_s^*\| \leq \|\mu_r-\nu_r^*\|+\|x-\mu_s\|+\|\mu_s-\nu_s^*\| \leq 2\frac{1}{\frac{1}{2\gamma}-2}\|x-\mu_s\|+\|x-\mu_s\| = \frac{1}{1-4\gamma}\|x-\mu_s\|$. $\square$

Proof of Lemma 2

Proof. Consider $G_{\max}$ obtained by adding all edges in E_{in}^* to G_0. Clearly, $G_{\max}$ has k connected components, where each component corresponds to a vertex set V_r^* for some $r \in k$. Adding any more edges from E_{out}^* to $G_{\max}$ will reduce the number of components to $k-1$. Furthermore, any $e \in E_{out}^*$ can only be added to G_{SL} after all edges in E_{in}^* are added. This means the algorithm must stop before any edges in E_{out}^* are added. This in turn implies the final solution G_{SL}, if not equal to $G_{\max}$, can be obtained by removing edges in $G_{\max}$. Since removing edges can only maintain or disconnect existing connected components and G_{SL} has the same number of connected components as that of $G_{\max}$, G_{SL} must have exactly the same k connected components as those of $G_{\max}$, so each component V_{SL}^r of G_{SL} corresponds to exactly one cluster V_r^* for some r. $\square$

Proof of Lemma 3

Proof. We first show without any assumption, if we sample X i.i.d. uniformly at random, then for each optimal cluster T_r, if $\nu_i \in T_r$, then $\|\nu_i-\mu_r\|$ satisfies the bound in A with high probability. Let $q := \|\nu_i-\mu_r\|^2$, we have $0 \leq q \leq \max_{x\in T_r}\|x-\mu_r\|^2$ and $E[q|\nu_i \in T_r] = \frac{\sum_{x\in T_r}\|x-\mu_r\|^2}{n_r} = \frac{\phi_*^r}{n_r}$. Then applying Hoeffding's bound, we get, $Pr\{q-Eq \geq (\frac{f}{4}-1)\frac{\phi_*^r}{n_r}|\nu_i \in T_r\} \leq \exp\{-\frac{2[(\frac{f}{4}-1)\frac{\phi_*^r}{n_r}]^2}{(\max_{x\in T_r}\|x-\mu_r\|^2)^2}\}$ Substituting $w_{\min}$ for every r and applying union bound, we get $Pr(A^c) \leq m\exp(-2(\frac{f}{4}-1)^2 w_{\min}^2)$. Now the probability of a cluster T_r not being seeded after m trials is $(1-p_r)^m \leq \exp(-mp_r)$. Applying union bound, we get $Pr(B^c) \leq k\exp(-mp_{\min})$. Applying union bound again, we get $Pr(A\cap B) \geq 1-m\exp(-2(\frac{f}{4}-1)^2w_{\min}^2)-k\exp(-mp_{\min})$. $\square$

Proof of Lemma 4

Proof. Let $\pi(i)=\pi(j)=r$. Then $\|\nu_i-\nu_j\| \leq \|\nu_i-\mu_r\|+\|\nu_j-\mu_r\| \leq 2\frac{\sqrt{f}}{2}\sqrt{\frac{\phi_*^r}{n_r}}$. Let $\pi(p)=t, \pi(q)=s$. Then $\|\nu_p-\nu_q\| \geq \|\mu_t-\mu_s\|-\|\nu_p-\mu_t\|-\|\nu_q-\mu_s\| \geq f\sqrt{\phi_1+\phi_2}(\frac{1}{\sqrt{n_t}}+\frac{1}{\sqrt{n_s}})-\frac{\sqrt{f}}{2}\sqrt{\frac{\phi_*^t}{n_t}}-\frac{\sqrt{f}}{2}\sqrt{\frac{\phi_*^s}{n_s}} > \frac{f}{2}\sqrt{\phi_1+\phi_2}(\frac{1}{\sqrt{n_t}}+\frac{1}{\sqrt{n_s}})$, by center-separability. On the other hand, recall $\alpha := \min_{r\neq s}\frac{n_r}{n_s}$, we get $\sqrt{\frac{1}{n_r}} \leq \min\{\frac{1}{\sqrt{\alpha n_t}}, \frac{1}{\sqrt{\alpha n_s}}\}$, so $2\sqrt{f}\sqrt{\frac{\phi_*^r}{n_r}} \leq \sqrt{f\phi_*^r}(\frac{1}{\sqrt{\alpha n_t}}+\frac{1}{\sqrt{\alpha n_s}})$. Since $f > \frac{1}{\alpha}$, we get $\|\nu_i-\nu_j\| \leq \sqrt{f}\sqrt{\frac{\phi_*^r}{n_r}} \leq \frac{f}{2}\sqrt{\phi_*^r}(\frac{1}{\sqrt{n_t}}+\frac{1}{\sqrt{n_s}}) < \frac{f}{2}\sqrt{\phi_1+\phi_2}(\frac{1}{\sqrt{n_t}}+\frac{1}{\sqrt{n_s}}) < \|\nu_p-\nu_q\|$. $\square$

Proof of Theorem 3

Proof. Consider $A\cap B$. Under this event, we know that the optimal clustering T_* induces a non-degenerate k-clustering of $\{\nu_i, i\in[m]\}$, which we denote by $\{V_r^*, r\in[k]\}$ with $V_r^* := T_r \cap \{\nu_i, i\in[m]\}, \forall r\in[k]$. In addition, Lemma 4 implies the bi-partite edge sets E_{in}^* and E_{out}^* induced by $\{V_r^*, r\in[k]\}$ satisfies $\forall e_1 \in E_{in}^*, e_2 \in E_{out}^*, w(e_1) < w(e_2)$. Thus, by Lemma 2, if we apply Single-Linkage on $G_0 = (\cup_{r\in[k]}V_r^*, \emptyset)$ until k components remain, each returned connected component $\tilde{S}_r$ corresponds to exactly one cluster V_r^*. In addition, with the seeding guarantee by event A, $\forall r\in[k], \|m(V_r^*)-\mu_r\| \leq \frac{1}{|V_r^*|}\sum_{\nu_i\in V_r^*}\|\nu_i-\mu_r\| \leq \frac{\sqrt{f}}{2}\sqrt{\frac{\phi_*^r}{n_r}}$. Noting $Pr(A\cap B) \geq 1-m\exp(-2(\frac{f}{4}-1)^2w_{\min}^2)-k\exp(-mp_{\min})$ by Lemma 3 and $m(V_r^*)=\nu_r^*$ completes the proof. $\square$

Theorem (Theorem 1 of [1]). *Assume there is a dataset-solution pair (X, T_*) satisfying $d_{rs}^*(f)$-center separability, with $f > 32$. If at iteration t, $\forall r\in[k], \Delta_r^t < \beta_t\frac{\sqrt{\phi_*}}{\sqrt{n_r}}$ with $\beta_t < \max\{\gamma\frac{f}{8}, \frac{128}{9f}\}$ with $\gamma < 1$, then $\forall r\in[k], \Delta_r^{t+1} < \beta_{t+1}\frac{\sqrt{\phi_*}}{\sqrt{n_r}}$, with $\beta_{t+1} < \max\{\frac{\gamma}{2}\frac{f}{8}, \frac{128}{9f}\}$.*

Theorem (Theorem 4 of [1]). *Assume* (X, T_*) *satisfies* $d^*_{rs}(f)$*-center separability with* $f > \frac{1}{\alpha}$. *If we obtain seeds* $\{\nu^*_r, r \in [k]\}$ *by applying the Heuristic clustering algorithm (Algorithm 1 here) to* X. *Then* $\forall \mu_r, \exists \nu^*_r$ *s.t.* $\|\mu_r - \nu^*_r\| \leq \frac{\sqrt{f}}{2}\sqrt{\frac{\phi^r_*}{n_r}}$ *with probability at least* $1 - m\exp(-2(\frac{f}{4}-1)^2 w^2_{\min}) - k\exp(-mp_{\min})$.

References

[1] Anonymous Authors. On Lloyd's algorithm: new theoretical insights for clustering in practice. Submitted, 2015.