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 $[\frac{1}{p_{\min}}\log\frac{2k}{\delta}, \frac{\delta}{2}\exp(2(\frac{f}{4}-1)^2 w_{\min}^2)]$. □

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 $\Omega(\frac{1}{\sqrt{\epsilon}})$. □

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 $Pr(A \cap B) \geq 1 - m\exp(-2(\frac{f}{4}-1)^2 w_{\min}^2) - k\exp(-mp_{\min})$ by Lemma 3 and $m(V_r^*) = \nu_r^*$ completes the proof. □

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* $1 - m\exp(-2(\frac{f}{4}-1)^2 w_{\min}^2) - k\exp(-mp_{\min})$.

References

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