1 Parallel k-Means in Theory and Practice

- Input:
  - Number of desired means $k \in \mathbb{N}$
  - Set of points $P \subset \mathbb{R}^d$ (or multiset)
- Output:
  - Set of k means $C = \{c_1, \ldots, c_k\}$

Notation

- $\mathrm{mean}(P) := \frac{1}{|P|} \cdot \sum_{x \in P} x$
- $\mathrm{dist}(x, c) := \|x - c\|^2$ and $\mathrm{dist}(x, C) := \min_{c \in C} \mathrm{dist}(x, c)$
- Could choose different functions mean and dist!
- Objective (potential) function: For $Q \subseteq P$, let $\Phi_Q(C) := \sum_{x \in Q} \mathrm{dist}(x, C)^2$
- For convenience, $\Phi := \Phi_P$

1.1 Lloyd's Heuristic

- Additional input: Set C of initial k clusters ("seeding").
- Simple (and unfortunately often-used) seeding strategy: k random points from P

1: **repeat**
2:     **for** $x \in P$ **do**
3:         $a[x] \leftarrow \arg\min_{c \in C} \mathrm{dist}(x, c)$
4:     **for** $c \in C$ **do**
5:         $c \leftarrow \mathrm{mean}(\{x \in P \mid a[x] = c\})$
6: **until** C did not change in last iteration

- **Theorem:** Lloyd's heuristic converges
- There are only finitely many point-to-centroid assignments $a[\cdot]$. In each step, Φ decreases: In step 3 by definition, and in step 5 by the fact that the mean minimizes the sum of squared distances in any single cluster:
- **Lemma:** Let $A \subset \mathbb{R}^d$, $c = \mathrm{mean}(A)$, $c' \in \mathbb{R}^d$ arbitrary. Then:

$$\sum_{x \in A} \|x - c'\|^2 = \sum_{x \in A} \|x - c\|^2 + |A| \cdot \|c' - c\|^2$$

- By definition (alternatively, just recall the law of cosine):

$$\begin{aligned} \|x - c'\|^2 &= \langle (x - c) - (c' - c), (x - c) - (c' - c) \rangle \\ &= \|x - c\|^2 - 2 \underbrace{\langle x - c, c' - c \rangle}_{\substack{\|x-c\| \cdot \|c'-c\| \\ \cdot \cos(\sphericalangle(x-c, c'-c))}} + \|c' - c\|^2 \end{aligned}$$

  Now $\sum_{x \in P} \langle x - c, c' - c \rangle = \langle \sum_{x \in P} (x - c), c' - c \rangle = 0$ due to the bilinearity of the dot product and definition of c.

1.2 k-Means++ Seeding

- Intuitively: When points P well separated, initial centroids should be from different clusters
- First idea: Choose initial centroids one-by-one, each time picking the furthest point. Remark: gives a 2-approximation for k-center problem: $\min_a \max_{x \in P} \operatorname{dist}(x, a[x])$
- Problem with k-means: Susceptible to outliers.
- Note: random and furthest-point strategies are at opposite ends of the same spectrum: Sample each new centroid with probability proportional to $\operatorname{dist}^\alpha(p, C)$. Random: $\alpha = 0$, furthest point: $\alpha = \infty$.

```
C ← {random p ∈ P}
while |C| < k do
    C ← C ∪ {random p ∈ P with probability proportional to dist(p, C)^2}
```

- **Theorem:** $\operatorname{E}[\Phi(C)] \leq 8(\log k + 2) \cdot \Phi(C^*)$
- Idea: 1. Show competitiveness for all clusters that k-means++ samples a point from. 2. More complicated proof necessary for clusters not "hit".
- **Lemma:** Let A be an arbitrary cluster in C^*. Let $C = \{\text{random } p \in A\}$. Then,

$$\begin{aligned} \operatorname{E}[\Phi_A(C)] &= \sum_{c' \in A} \frac{1}{|A|} \cdot \sum_{x \in A} \|x - c'\|^2 \\ &= \sum_{c' \in A} \frac{1}{|A|} \cdot \left(\sum_{x \in A} \|x - c\|^2 + |A| \cdot \|c' - c\|^2 \right) \\ &= 2 \sum_{x \in A} \|x - c\|^2 = 2 \cdot \Phi_A(C^*). \end{aligned}$$

- **Lemma:** Let A be an arbitrary cluster in C^*, and C be an arbitrary clustering. Let $C' = C \cup \{\text{random } p \in A \text{ with probability proportional to } \operatorname{dist}(p, C)^2\}$. Then, $\operatorname{E}[\Phi_A(C')] \leq 8\Phi_A(C^*)$.
- We have:

$$\operatorname{E}[\Phi_A(C')] = \sum_{c' \in A} \frac{\operatorname{dist}(c', C)^2}{\sum_{x \in A} \operatorname{dist}(x, C)^2} \cdot \sum_{x \in A} \min(\operatorname{dist}(x, C), \|x - c'\|)^2$$

- Triangle inequality: $\operatorname{dist}(c', C) \leq \|c' - x\| + \operatorname{dist}(c, C)$
- Cauchy-Schwarz: $\operatorname{dist}(c', C)^2 \leq 2 \cdot \|c' - x\|^2 + 2 \cdot \operatorname{dist}(c, C)^2$
- Summing up over all $x \in A$: $\operatorname{dist}(c', C)^2 \leq \frac{2}{|A|} \sum_{x \in A} \operatorname{dist}(x, C)^2 + \frac{2}{|A|} \sum_{x \in A} \|x - c'\|^2$
- Putting everything together:

$$\begin{aligned} \operatorname{E}[\Phi_A(C')] &= \sum_{c' \in A} \left[\frac{2}{|A|} \sum_{x \in A} \|x - c'\|^2 + \frac{2}{|A|} \sum_{x \in A} \|x - c'\|^2 \right] = 4 \cdot \sum_{c' \in A} \frac{1}{|A|} \cdot \sum_{x \in A} \|x - c'\|^2 \\ &= 8\Phi_A(C^*) \qquad \text{[previous Lemma]} \end{aligned}$$

- **Lemma:** (about $C, B_1, \ldots, B_u, t$) Let C be an arbitrary clustering, X be the set of points that are in clusters of C^* hit by C, and $B_1, \ldots, B_u$ be clusters in C^* not hit by C. Define $U = \bigcup_{i=1}^{u} B_i$. Suppose we add $t \le u$ random centers to C, as in line 3. Then

$$\mathrm{E}[\Phi_{X \cup U}(C')] \le (\Phi_X(C) + 8 \cdot \Phi_U(C^*)) \cdot (1 + H_t) + \frac{u-t}{u} \cdot \Phi_U(C)$$

- Proof by induction over (t, u). Two base cases:
  i) $t = 0, u > 0$
  ii) $t = 1, u = 1$

- Induction step: Prove that, if the hypothesis holds for $(t-1, u)$ and $(t-1, u-1)$, then it also holds for (t, u).

- Consider case (t, u): Denote by c' the first center added to C. Two cases for which to compute conditional expectation:
  i) $c' \in X$: Invoke IH with

$$C \cup \{c'\}, (B_1, \ldots, B_u), t-1$$

  Conditional expectation:

$$\mathrm{E}[\Phi_{X \cup U}(C') \mid c' \in X] \le (\Phi_X(C) + 8 \cdot \Phi_U(C^*)) \cdot (1 + H_{t-1}) + \frac{u-t}{u} \cdot \Phi_U(C) + \underbrace{\frac{1}{u} \cdot \Phi_U(C)}_{\le \frac{1}{t} \cdot \Phi_{X \cup U}(C)}$$

  The last term is the reason for H_t appearing in the non-conditional expectation!

  ii) $c' \notin X$: Hence, there is an i with $c' \in B_i$. For each B_i, invoke IH with:

$$C \cup \{c'\}, (B_1, \ldots, B_{i-1}, B_{i+1}, \ldots, B_u), t-1$$

  Sum up to get conditional expectation:

$$\mathrm{E}[\Phi_{X \cup U}(C') \mid c' \notin X] \le (\Phi_X(C) + 8 \cdot \Phi_U(C^*)) \cdot (1 + H_{t-1}) + \frac{u-t}{u} \cdot \Phi_U(C)$$

  Obviously, case (i) has probability $\frac{\Phi_X(C)}{\Phi_{X \cup U}(C)}$, and case (ii) has the complementary probability.

- The math is relatively straightforward, though it does involve a few tricks (e.g., using Cauchy-Schwarz again). Of course, the previous lemma has to be used as well.

- **Proof of Theorem:** Consider C after line 1. Let $B_1, \ldots, B_{k-1}$ be the clusters in C^* not hit by C. Invoke the previous lemma with $C, (B_1, \ldots, B_{k-1}), k-1$. Note that $P = X \cup U$ (notation as in the previous lemma).

$$\mathrm{E}[\Phi(C') \mid C] \le (\Phi_X(C) + 8 \cdot \underbrace{\Phi_U(C^*)}_{= \Phi(C^*) - \Phi_X(C^*)}) \cdot (1 + \underbrace{H_{k-1}}_{\le 1 + \ln k})$$

  The claim follows because $\mathrm{E}[\Phi_X(C)] \le 2 \cdot \Phi_X(C^*)$ by the first lemma.

1.3 k-means||

- Problem: k-means++ is inherently sequential
- Again: Random sampling and k-means++ can be seen as the two ends on the spectrum: Sample all k centers in one iteration vs. sample one center in each of k iterations (distribution depends on previous iterations)

```
1: C ← {random p ∈ P}
2: Φ_0 ← Φ(C)
3: for O(log Φ_0) times do
4:     C′ ← {sample each p ∈ P independently with probability ℓ·dist(p,C)^2 / Φ(C)}
5:     C ← C ∪ C′
6: for c ∈ C do
7:     w_c ← number of points in P that are closer to c than to any other point in C
8: Run (weighted) k-means++ on C
```

- **Theorem** (no proof): Before line 8, $\Phi(C) = O(\Phi(C^*))$. (Note that C has $O(\ell \log \Phi_0)$ centroids.)
- **Lemma** (no proof): Let C be a (fixed) set of centroids. After executing line 4, we have $\mathrm{E}[\Phi(C \cup C')] \leq 8\Phi(C^*) + \alpha \cdot \Phi(C)$, where $\alpha \in (0, 1)$ only depends on ℓ and k.
- **Corollary:** Let $C = \{p\}$. Denote by C^i the the (random) value of C at the end of iteration i. Then:

$$\mathrm{E}[\Phi(C^i)] \leq \alpha^i \cdot \Phi_0 + \frac{8}{1-\alpha}\Phi(C^*)$$

- Base case: $i = 0$ is trivial.
- Induction step: By theorem:

$$\mathrm{E}[\Phi(C^{i+1} \mid C^i] \leq \alpha \cdot \Phi(C^i) + 8\Phi(C^*)$$

  Can take expectation over C^i:

$$\begin{aligned}\mathrm{E}[\Phi(C^{i+1})] &\leq \alpha \cdot \mathrm{E}[\Phi(C^i)] + 8\Phi(C^*) \\ &= \alpha \cdot \left(\alpha^i \Phi_0 + \frac{8}{1-\alpha}\Phi(C^*)\right) + 8\Phi(C^*) \\ &= \alpha^{i+1} \cdot \Phi_0 + \frac{8}{1-\alpha} \cdot \Phi(C^*)\end{aligned}$$

- Now if $i = -\log_\alpha \Phi_0$, we have $\alpha^i \cdot \Phi_0 = 1$, i.e., $\mathrm{E}[\Phi(C^i)] = O(\Phi(C^*))$.

1.4 k-Means on a (Hemi-)Sphere

- Commonly used metric for k-means on text data: Angles between feature vector. For instance term-frequency/inverse-document-frequency (tf-idf). Let D be a set of documents, T be a set of terms ("dictionary"), $\mathrm{tf}(t, d)$ denote the number of occurrences of term t in document $d \in D$, and $\mathrm{idf}(t) = \log \frac{|D|}{|\{d \in D | t \in D\}|}$. Represent each document $d \in D$ as vector:

$$\big(\mathrm{tf}(t, d) \cdot \mathrm{idf}(t)\big)_{t \in T}$$

- Typical metric used is the angle between two documents (sometimes called "cosine similarity"). Conceptually, we can think of each document as a point on the sphere $S^{|T|-1}$.
- Idea: Cluster according to topic, not length! Roughly: A document concatenated with itself should have distance 0 from the original.
- MADlib v0.4 for k-means with "cosine" metric:
  - Closest centroid: Choose smallest angle
  - Mean of points: Normalized Euclidean mean
- Problem: Spherical mean (i.e., w.r.t. geodesic distances) and normalized Euclidean mean do not coincide in general.

  Example on S^1: Let there be a points at $(1,0)$ and b points at $(0,1)$. Angle between x-axis and spherical average should be $\frac{a}{a+b} \cdot \frac{\pi}{2}$. Using Euclidian mean:
  - mean of the $a+b$ points is $(\frac{a}{a+b}, \frac{b}{a+b})$
  - Angle between x-axis and Euclidean mean is $\arctan(\frac{a}{b})$.

  Substitute $\alpha = \frac{a}{b}$: Clearly, $\arctan(\alpha)$ and $\frac{\alpha}{\alpha+1} \cdot \frac{\pi}{2}$ are not identical.
- Does k-means converge at all? No approximation guarantees for k-means phase!
- Good news: Convergence guaranteed when using "Euclidean" objective as potential function. Also approximation guarantees for this potential.
- Alternative: Use spherical average: Must minimize sum of squared distanced.
- **Lemma:** Let $A \subset S^d$ finite, $\gamma = \| \operatorname{mean}(A) \|$, $c = \frac{\operatorname{mean}(A)}{\gamma}$, $c' \in S^d$ arbitrary. Then:

$$\sum_{x \in A} \|x - c\|^2 \leq \sum_{x \in A} \|x - c'\|^2 \leq \sum_{x \in A} \|x - c\|^2 + |A| \cdot \|c' - c\|^2$$

- Like at the beginning:

$$\|x - c'\|^2 = \|x - c\|^2 + \langle 2 \cdot (c - x), c' - c \rangle + \|c' - c\|^2$$

  Here:

$$\begin{aligned}
\sum_{x \in P} \langle 2 \cdot (c - x), c' - c \rangle &= \left\langle 2 \cdot \sum_{x \in P} (c - x), c' - c \right\rangle \\
&= \left\langle 2 \cdot (|P| \cdot c - |P| \cdot \gamma \cdot c), c' - c \right\rangle \\
&= \langle 2 \cdot |P| \cdot (1 - \gamma) \cdot c, c' - c \rangle \qquad (1.1)
\end{aligned}$$

  Now for the upper bound, note that (1.1) is the same as:

$$2 \cdot |P| \cdot \underbrace{(1 - \gamma)}_{\geq 0} \cdot \underbrace{(\langle c, c' \rangle - \underbrace{\langle c, c \rangle}_{=1})}_{\leq 0} \leq 0$$

For the lower bound, not that (1.1) plus $|P| \cdot \|c - c'\|$ is the same as:

$$\begin{aligned} & |P| \cdot \langle 2 \cdot (1-\gamma) \cdot c + c' - c, c' - c \rangle \\ = & |P| \cdot \langle c' - (2\gamma - 1) \cdot c, c' - c \rangle \end{aligned}$$

W.l.o.g. (rotate all points), we can assume that $c = (0, \ldots, 0, 1)$. Then, the previous is equal to:

$$\begin{aligned} & |P| \cdot \left[\sum_{i=1}^{d-1} c_i'^2 + \underbrace{(c_d' - (2\gamma - 1))(c_d' - 1)}_{c_d'^2 - c_d' - (2\gamma-1) \cdot c_d' + (2\gamma - 1)} \right] \\ = & |P| \cdot [1 - 2\gamma \cdot c_d' + 2\gamma - 1] \\ = & |P| \cdot (1 - c_d') \geq 0 \end{aligned}$$