Data Science Through a Geometric Lens Fall 2019

Lecture 11 & 12 — November 6 & November 13 2019

Prof. Cyrus Rashtchian *Topics: Clustering: k-means*

Overview. Last time we talked about clustering in general, and its many flavors. And we also presented a 2-approximation for k-center clustering in any metric space. Today we will talk about k-means clustering. Next lecture, we will prove that k-means++ provides an $O(\log k)$ approximation to the optimal k-means solution [1]. The notes for both lectures are combined in this document.

1 The k-means problem

Recall that we denote $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ as the points in $\mathcal{X}$, and we define the n-dimensional distance vector

$$\mathcal{X}_T = [d(x_1, T), d(x_2, T), \ldots, d(x_n, T)].$$

Given a dataset $\mathcal{X}$ in a metric space, the objective is to find a set T with size k that minimizes

$$\text{cost}(T) = \|\mathcal{X}_T\|_2^2.$$

The k-means cost function can equivalently be written as

$$\text{cost}(T) = \|\mathcal{X}_T\|_2^2 = \sum_{x \in \mathcal{X}} \min_{c \in T} \|x - c\|_2^2,$$

where intuitively we are paying for the distance from x to the closest center $c \in T$.

In this lecture, we will focus on input sets $\mathcal{X} \subseteq \mathbb{R}^d$ being a subset of real vectors, with the ℓ_2 distance. We also consider the possibility of T not necessarily being a subset of $\mathcal{X}$. Formally, we have the following.

Definition 1 (k-means clustering)**.** *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a dataset. Find a set of k points $T \subseteq \mathbb{R}^d$ that minimizes*

$$\text{cost}(T) = \|\mathcal{X}_T\|_2^2 = \sum_{x \in \mathcal{X}} \min_{c \in T} \|x - c\|_2^2,$$

1.1 The simple case of $k = 1$

To gain some intuition, consider finding one center c. Then, the best thing to do is to take the mean of the data. We prove a robust version of this claim in Lemma 2. To this end, we define for any set $C \subseteq \mathbb{R}^d$,

$$\text{mean}(C) = \frac{1}{|C|} \cdot \sum_{y \in C} y.$$

When $k = 1$, then setting $T = \{\mu\}$ as the single cluster center, for the mean $\mu = \text{mean}(\mathcal{X})$, will minimize the 1-means cost. Analogously, if we wanted to find a single 'representative' for a cluster, then the best thing to do is to take the mean of the points in the cluster. It will be convenient to define the cost of a single cluster and a single center as follows:

$$\text{cost}(C, z) = \sum_{x \in C} \|x - z\|_2^2.$$

We can see exactly how much the cost changes if we choose a center other than the mean.

Lemma 2. *For any $C \subseteq \mathbb{R}^d$, let $\mu = \text{mean}(C)$. For any $z \in \mathbb{R}^d$, we have*

$$\text{cost}(C, z) = \text{cost}(C, \mu) + |C| \cdot \|z - \mu\|_2^2.$$

This lemma follows from a more general random variable fact. The main takeaway is that the squared ℓ_2 norm is nice to work with when it comes to sums/expectations.

Lemma 3. *Let $Y \in \mathbb{R}^d$ be a random vector. For any $z \in \mathbb{R}^d$, we have*

$$\mathbb{E}\,\|Y - z\|_2^2 = \mathbb{E}\,\|Y - \mathbb{E}\,Y\|_2^2 + \|z - \mathbb{E}\,Y\|_2^2$$

Proof. Let $\mu = \mathbb{E}\,Y \in \mathbb{R}^d$. Expanding the right hand side we have

$$\begin{aligned}
\mathbb{E}\,\|Y - \mu\|_2^2 + \|z - \mu\|_2^2 &= \mathbb{E}\left[\|Y\|_2^2 + \|\mu\|_2^2 - 2\langle Y, \mu\rangle\right] + \|z\|_2^2 + \|\mu\|_2^2 - 2\langle z, \mu\rangle \\
&= \mathbb{E}\,\|Y\|_2^2 + \|\mu\|_2^2 - 2\langle \mu, \mu\rangle + \|z\|_2^2 + \|\mu\|_2^2 - 2\langle z, \mu\rangle \qquad \text{(linearity of expectation)} \\
&= \mathbb{E}\,\|Y\|_2^2 + \|z\|_2^2 - 2\langle z, \mu\rangle \\
&= \mathbb{E}\,\|Y - z\|_2^2.
\end{aligned}$$

□

Then we can prove the previous lemma as follows:

Proof of Lemma 2. Let Y denote a uniformly random point in C. Then,

$$\mathbb{E}\,\|Y - z\|_2^2 = \frac{1}{|C|} \cdot \sum_{y \in C} \|y - z\|_2^2 = \frac{1}{|C|} \cdot \text{cost}(C, z).$$

Also,

$$\mathbb{E}\,\|Y - \mu\|_2^2 = \frac{1}{|C|} \cdot \text{cost}(C, \mu).$$

Applying Lemma 3 proves Lemma 2 after multiplying by $|C|$,

$$\frac{1}{|C|} \cdot \text{cost}(C, z) = \frac{1}{|C|} \cdot \text{cost}(C, \mu) + \|z - \mu\|_2^2.$$

□

This is useful to understand what happens if we use a random cluster point as the center.

Lemma 4. *For any $C \subseteq \mathbb{R}^d$, let $\mu = \text{mean}(C)$. If z is chosen uniformly at random from C, then*

$$\mathbb{E}[\text{cost}(C, z)] = 2 \cdot \text{cost}(C, \mu).$$

Proof. We use Lemma 2. Let $\mu = \text{mean}(C)$. Then,

$$\begin{aligned}
\mathbb{E}[\text{cost}(C, z)] &= \frac{1}{|C|} \cdot \sum_{z \in C} \text{cost}(C, z) \\
&= \frac{1}{|C|} \cdot \sum_{z \in C} \left(\text{cost}(C, \mu) + |C| \cdot \|z - \mu\|_2^2\right) \\
&= \text{cost}(C, \mu) + \sum_{z \in C} \|z - \mu\|_2^2 \\
&= 2 \cdot \text{cost}(C, \mu).
\end{aligned}$$

□

In other words, if $k = 1$, then picking a random point in $\mathcal{X}$ is already a 2-approximation.

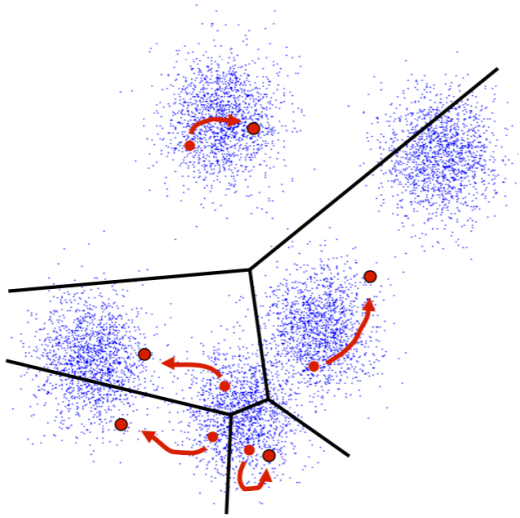

Figure 1: One iteration of the k-means, where the centers change to the new means of the clusters. The black lines correspond to the Voronoi diagram of the previous centers.

2 k-means algorithm

In 1957 Stuart Lloyd suggested a simple iterative algorithm which efficiently finds a local minimum for this problem. This algorithm (a.k.a. Lloyd's algorithm) seems to work so well in practice that it is sometimes referred to as k-means or the k-means algorithm.

The k-means algorithm

`Initialize` $c_1, c_2, \ldots, c_k \in \mathbb{R}^d$ `and define clusters` $C_1, C_2, \ldots, C_k$ `arbitrarily`
`While something changes :`
    `for each` $j \in [k]$: $C_j \leftarrow \{x \in \mathcal{X} \mid c_j$ `is the closest center to` $x\}$
    `for each` $j \in [k]$: $c_j = \text{mean}(C_j)$
`Output` $T = \{c_1, c_2, \ldots, c_k\}$.

The time per iteration is $O(kn)$, and the algorithm always converges to some set of k centers.

Lemma 5. *The k-means algorithm always converges in a finite number of steps.*

Proof. We show that the cost monotonically decreases. Let $c_1^{(t)}, c_2^{(t)}, \ldots, c_k^{(t)}$ be the k centers at time t, and similarly, let $C_1^{(t)}, C_2^{(t)}, \ldots, C_k^{(t)}$ be the k clusters at time t (where at time t, we mean at the start of the t^{th} iteration). The first step of the while loop assigns each point to the closest center. Therefore,

$$\text{cost}(C_{1:k}^{(t+1)}, c_{1:k}^{(t)}) \leq \text{cost}(C_{1:k}^{(t)}, c_{1:k}^{(t)}).$$

In second step of the while loop,

$$\text{cost}(C_{1:k}^{(t+1)}, c_{1:k}^{(t+1)}) \leq \text{cost}(C_{1:k}^{(t+1)}, c_{1:k}^{(t)}).$$

So, the cost can never increase. Moreover, if any of the centers move, then the cost will strictly decrease because the mean of the clusters determines the best centers for those clusters. □

However, the initialization is very important. For most initial centers $c_1, \ldots, c_k$, the issue is that the algorithm above will converge to a local optima that is not very good.

3 k-means++ initialization

`Choose` $z \in \mathcal{X}$ `uniformly at random and initialize` $T_1 = \{z\}$

`For` $i = 1, 2, 3, \ldots, k-1$:

  `Choose` z `with probability proportional to` $d(z, T_i) = \min_{c \in T_i} \|x - c\|_2^2$

  $T_{i+1} \leftarrow T_i \cup \{z\}$

`Output` $T \leftarrow T_k$.

This initialization takes time $O(kn)$, which is the same as a single iteration of the k-means algorithm. We will show that in expectation, this initialization already gives a pretty good clustering. Moreover, iterating the algorithm afterwards will improve the cost (and hence is useful in practice).

The main result we will prove about k-means is the following.

Theorem 6. *Let T be the k centers chosen by k-means++, and let T^* be the optimal k centers. Then*

$$\mathbb{E}[\text{cost}(T)] \leq \text{cost}(T^*) \cdot O(\log k),$$

where the expectation is over the randomness in the initialization procedure.

We need a handful of lemmas to prove this theorem.

3.1 Analysis

Let $T^* = \{c_1^*, c_2^*, \ldots, c_k^*\}$ denote the optimal k-means centers, and let $C_1^*, C_2^*, \ldots, C_k^*$ denote the corresponding clusters. Notice that this implies that $c_j^* = \text{mean}(C_j^*)$.

The first center we choose is a uniformly random point in $\mathcal{X}$. If it lands in a cluster C_j^*, then it is a random point from C_j^*. Hence, in expectation, the cost is at most twice the optimal based on the optimal center c_j^* by Lemma 4.

The challenge is that Lemma 4 does not apply to subsequent centers, because they are not uniform draws from any cluster. In particular, the next points are farther away from the already chosen centers.

We next show that if we hit another cluster later on, then the cost is at most eight times more than the true cost. The overall analysis will use this to our advantage, and the main work will be in proving the "uncovered" clusters do not contribute to much to the cost of the returned clustering.

Lemma 7. *If some centers T_i have already been chosen by k-means++ and $z \in C_j$ is added next, then*

$$\mathop{\mathbb{E}}_{z}[\text{cost}(C_j, z) \mid T_i, z \in C_j] \leq 8 \cdot \text{cost}(C_j, c_j^*).$$

Proof. We consider $T_{i+1} = T_i \cup \{z\}$. For any x, we have $\text{cost}(x, T_{i+1}) = \min\{cost(x, T_i),\ \|x - z\|_2^2\}$.

We start by writing out the LHS in the lemma statement in terms of the conditioning:

$$\begin{aligned}
\mathop{\mathbb{E}}_{z}[\text{cost}(C_j, z) \mid T_i, z \in C_j] &= \sum_{z \in C_j} \Pr[\text{algorithm chooses } z \mid T_i]\text{cost}(C_j, T_{i+1}) \\
&= \sum_{z \in C_j} \frac{\text{cost}(z, T_i)}{\text{cost}(C_j, T_i)}\text{cost}(C_j, T_{i+1}) \\
&= \sum_{z \in C_j} \frac{\text{cost}(z, T_i)}{\text{cost}(C_j, T_i)} \sum_{x \in C_j} \min\{\text{cost}(x, T_i),\ \|x - z\|_2^2\}.
\end{aligned}$$

We will break this up into two parts. First, we upper bound $\text{cost}(z, T_i)$ by using the whole cluster C_j. For any $x \in C_j$, let c be its closest center in T_i. By the triangle inequality we have that

$$\begin{aligned}
\text{cost}(z, T_i) &\leq \|z - c\|_2^2 \\
&\leq (\|z - x\|_2 + \|x - c\|_2)^2 \\
&\leq 2\|z - x\|_2^2 + 2\|x - c\|_2^2 \\
&= 2\|z - x\|_2^2 + 2\text{cost}(x, T_i),
\end{aligned}$$

where the final inequality uses $(a+b)^2 \leq 2a^2 + 2b^2$ for any non-negative $a, b \in \mathbb{R}$, which follows from AMGM. Summing over $x \in C_j$, we have that

$$|C_j| \cdot \text{cost}(z, T_i) \leq 2 \sum_{x \in C_j} \|x - z\|_2^2 + 2 \sum_{x \in C_j} \text{cost}(x, T_i) = 2\text{cost}(z, C_j) + 2\text{cost}(C_j, T_i).$$

We will handle the factor of $|C_j|$ by preemptively multiplying by it in the following. Now we plug this into our rewriting of the LHS of the expectation:

$$\begin{aligned}
|C_j| \mathop{\mathbb{E}}_{z}[\text{cost}(C_j, z) \mid T_i, z \in C_j] &\leq \sum_{z \in C_j} \frac{|C_j|\text{cost}(z, T_i)}{\text{cost}(C_j, T_i)} \sum_{x \in C_j} \min\{\text{cost}(x, T_i),\ \|x - z\|_2^2\}. && (1) \\
&\leq \sum_{z \in C_j} \frac{2\text{cost}(z, C_j) + 2\text{cost}(C_j, T_i)}{\text{cost}(C_j, T_i)} \sum_{x \in C_j} \min\{\text{cost}(x, T_i),\ \|x - z\|_2^2\} && (2) \\
&\leq \left(\sum_{z \in C_j} \frac{2\text{cost}(z, C_j)}{\text{cost}(C_j, T_i)} \sum_{x \in C_j} \text{cost}(x, T_i)\right) + \left(\sum_{z \in C_j} 2 \sum_{x \in C_j} \|x - z\|_2^2\right) && (3) \\
&= \left(\sum_{z \in C_j} \frac{2\text{cost}(z, C_j)}{\text{cost}(C_j, T_i)} \sum_{x \in C_j} \text{cost}(x, T_i)\right) + \left(\sum_{z \in C_j} 2\text{cost}(z, C_j)\right) && (4) \\
&= \left(\sum_{z \in C_j} 2\text{cost}(z, C_j)\right) + \left(\sum_{z \in C_j} 2\text{cost}(z, C_j)\right) && (5) \\
&= 4 \sum_{z \in C_j} \text{cost}(z, C_j) && (6)
\end{aligned}$$

Next we can use Lemma 4 to rewrite this in terms of the optimal mean $c_j^* = \text{mean}(C_j)$,

$$4 \sum_{z \in C_j} \text{cost}(z, C_j) = 8|C_j| \cdot \text{cost}(C, c_j^*).$$

Diving by $|C_j|$, since we preemptively multiplied by it, finishes the proof. □

Combining these two lemmas we have that the first center is a 2-approximation for its respective cluster. Then, the subsequent centers are an 8-approximation for their clusters, if they are in a cluster. But, the overall approximation will be an $O(\log k)$ approximation because we may fail to choose points in certain clusters.

3.2 Main Analysis

We need quite a bit of notation to keep track of several random variables that are relevant to the execution of the algorithm over time. The main idea is to keep track of all the additions of new centers to the eventual output. Each time, there will be a good case and a bad case. But we will show that in expectation there is a good enough balance to get the approximation we desire. Overall, the argument is somewhat subtle, but each of the steps will be fairly easy.

Define the following. Let $i = 0, 1, \ldots, k$ be the steps of the k-means++ algorithm, where at time i, we have i centers in the solution.

- T_i = the set of centers chosen so far (where $|T_i| = i$).
- H_i = the set of optimal clusters that we have 'hit' by choosing a point in that cluster to be a center in T_i.
- $U_i = [k] \setminus H_i$ = the set of optimal clusters that 'unhit' because we haven't chosen any points in that cluster yet.
- $W_i = i - |H_i|$ = the number of 'wasted' iterations because we didn't hit a new cluster.
- For a set of (optimal) cluster indices $J \subseteq [k]$, we define

$$\text{cost}(J, T_i) = \sum_{j \in J} \sum_{x \in C_j^*} \|x - T_i(x)\|_2^2$$

  where $T_i(x)$ denotes the closest point to x in T_i, that is,

$$T_i(x) = \arg\min_{c \in T_i} \|x - c\|_2^2,$$

  which is the center from T_i for the (algorithm's) cluster containing x. This is consistent with the single-cluster notation that we had before because

$$\text{cost}(\{j\}, \{z\}) = \text{cost}(C_j^*, z) = \sum_{x \in C_j^*} \|x - z\|_2^2.$$

  Also, $\text{cost}(T^*) = \text{cost}([k], T^*)$ as this sums over all clusters, which partition the dataset $\mathcal{X}$.

- We use $\text{cost}_i(J) = \text{cost}(J, T_i)$ as shorthand for the cost of the (optimal) clusters indexed by $J \subseteq [k]$ when using the centers in T_i.

At the very beginning, when $i = 0$, we have $H_0 = \emptyset$ and $W_0 = 0$. At the very end, we will have $W_k = |U_k|$ because this is the number of centers we have missed in the k iterations. Therefore, we have that

$$\text{cost}_k(H_k) + \text{cost}_k(U_k) = \text{cost}_k(T_k).$$

The key idea is to set up a iterative charging scheme, so that we pay a little bit over time. In particular, we will consider

$$\text{cost}_i(H_i) + \frac{W_i \text{cost}_i(U_i)}{|U_i|}.$$

As we just observed, this is $\text{cost}_k(T_k)$ when $i = k$. But it will be useful to analyze the intermediate values as well.

First, notice that the H_i part we have already handled because we can show we get an 8-approximation for this part by using Lemma 7.

Lemma 8. *For all $i \leq k$ we have*

$$\mathbb{E}[\text{cost}_i(H_i)] \leq 8\text{cost}(T^*).$$

Therefore, we focus on the second term in our charging scheme, which we denote as

$$\Phi_i = \frac{W_i \text{cost}_i(U_i)}{|U_i|}.$$

Notice that by the definition of $W_i = i - |H_i|$ and $|U_i| = k - |H_i|$, we have that $W_i \leq |U_i|$ because $i \leq k$, and therefore, $\Phi_i \leq \text{cost}_i(U_i)$. We will prove that we pay a certain fraction of $\text{cost}_i(U_i)$ at each iteration, where the total will be at most $O(\text{cost}(T^*) \cdot \log k)$.

This is part of the reason that we use the phrase charging scheme. Another reason is that we will consider $\mathbb{E}[\Phi_{i+1} - \Phi_i]$, and we will eventually end up using a telescoping sum to get $\mathbb{E}[\Phi_k]$.

There are two core lemmas in the analysis, depending on whether iteration $i + 1$ hits a new cluster or not. It will be convenient to let $\mathcal{E}_i$ denote the random events of the algorithm up to and including time i. As usual, we let z denote the cluster center that we choose in iteration $i + 1$.

Hitting a new cluster

The good case. Intuitively, hitting a new cluster is going to be good for us, because. it leads to no additional charge in expectation. In this case we hit a new center, and the number of wasted iterations stays the same. We can formally show that the charge is non-positive as follows (note that j is a random variable).

Lemma 9. *If $z \in C_j$ and $j \in U_i$, then*

$$\mathbb{E}[\Phi_{i+1} - \Phi_i \mid \mathcal{E}_i, j \in U_i] \leq 0.$$

Proof. When $j \in U_i$, we have that $H_{i+1} = H_i \cup \{j\}$ and $W_{i+1} = W_i$ and $U_{i+1} = U_i \setminus \{j\}$. Then,

$$\Phi_{i+1} = \frac{W_{i+1}\text{cost}_{i+1}(U_{i+1})}{|U_{i+1}|} = \frac{W_i \text{cost}_{i+1}(U_i \setminus \{j\})}{|U_i| - 1} \leq \frac{W_i(\text{cost}_i(U_i) - \text{cost}_i(C_j))}{|U_i| - 1},$$

where the inequality used that z is some element in C_j.

In the k-means++ algorithm, a center in cluster C_j is chosen with probability $\text{cost}_i(C_j)/\text{cost}_i(U_i)$ since we are conditioning at the new center being from an unhit cluster.

We want to lower bound the expectation for Φ_{i+1} by the one for Φ_i. Therefore,

$$\mathbb{E}[\text{cost}_i(C_j) \mid \mathcal{E}_i,\ j \in U_i] = \sum_{\ell \in U_i} \text{cost}_i(C_\ell) \cdot \frac{\text{cost}_i(C_\ell)}{\text{cost}_i(U_i)}.$$

Using Cauchy-Schwarz, like from a few lectures ago (or because it's smallest when they all the numerators are equal), we can lower bound this by $\frac{\text{cost}_i(U_i)}{|U_i|}$.

Now we can put these things together and we have

$$\begin{aligned}\mathbb{E}[\Phi_{i+1} \mid \mathcal{E}_i,\ j \in U_i] &\leq \frac{W_i}{|U_i|-1} \cdot (\text{cost}_i(U_i) - \mathbb{E}[\text{cost}_i(C_j) \mid \mathcal{E}_i,\ j \in U_i]) \\ &\leq \frac{W_i}{|U_i|-1} \cdot \left(\text{cost}_i(U_i) - \frac{\text{cost}_i(U_i)}{|U_i|}\right) \\ &= \Phi_i.\end{aligned}$$

The final equality uses that $\frac{1}{a-1} - \frac{1}{a(a-1)} = \frac{1}{a}$ for any integer $a \geq 2$. Note that it's okay to assume $|U_i| \geq 2$, because this is the case where step $i+1$ hits a new cluster (and if at any point we hit all clusters then we would be done because we would get an 8-approx). □

Hitting an already hit cluster

The bad case. If we hit the same cluster twice, then we are going to be missing some cluster down the road. We analyze the incremental effect of this as follows (note that we will bound random variables by other random variables, without an expectation). Hitting the same cluster twice means W_i goes up by one because we wasted an iteration (and H_i and U_i stay the same).

Lemma 10. *If $j \in U_i$, then*

$$\Phi_{i+1} - \Phi_i \leq \frac{\text{cost}_i(U_i)}{|U_i|}.$$

Proof. When $j \in H_i$, we have that $H_{i+1} = H_i$ and $W_{i+1} = W_i + 1$ and $U_{i+1} = U_i$. Then,

$$\Phi_{i+1} - \Phi_i = \frac{W_{i+1}\text{cost}_{i+1}(U_{i+1})}{|U_{i+1}|} - \frac{W_i\text{cost}_i(U_i)}{|U_i|} = \frac{(W_i+1)\text{cost}_{i+1}(U_i)}{|U_i|} - \frac{W_i\text{cost}_i(U_i)}{|U_i|} \leq \frac{\text{cost}_i(U_i)}{|U_i|},$$

where the final inequality used that $\text{cost}_{i+1}(U_i) \leq \text{cost}_i(U_i)$ because more centers can only decrease the cost. □

Putting everything together

Lemma 11.

$$\mathbb{E}[\Phi_{i+1} - \Phi_i \mid \mathcal{E}_i] \leq \frac{\text{cost}_i(H_i)}{k-i}.$$

Proof. We have considered two cases. In the first, the incremental charge is non-positive, so it won't matter. In the second, we have a probability of

$$\frac{\text{cost}_i(H_i)}{\text{cost}(T_i)}$$

of hitting a cluster that was already hit. Hence, by the previous Lemma, this has expected cost

$$\mathbb{E}[\Phi_{i+1} - \Phi_i \mid \mathcal{E}_i] \leq \frac{\text{cost}_i(H_i)}{\text{cost}(T_i)} \cdot \frac{\text{cost}_i(U_i)}{|U_i|} \leq \frac{\text{cost}_i(H_i)}{|U_i|},$$

where we used that $\text{cost}(T_i) \geq \text{cost}_i(U_i)$ because the former sums over all the points in the dataset, whereas the latter only sums over U_i. Finally, since $|U_i| \geq k - i$ because $|H_i| \leq i$ and $U_i = [k] \setminus H_i$, we are done. □

Theorem 12. *If T_k are the k centers returned by k-means++, and T^* are the optimal centers, then*

$$\mathbb{E}[\text{cost}(T_k)] \leq 8(2 + \ln k) \cdot \text{cost}(T^*).$$

Proof. We have already seen that

$$\text{cost}(T_k) = \text{cost}_k(H_k) + \text{cost}_k(U_k) = \text{cost}_k(H_k) + \Phi_k.$$

It is also possible to show (or just take it on faith) that

$$\mathbb{E}[\text{cost}(T_k)] = \mathbb{E}[\text{cost}_k(H_k)] + \mathbb{E}[\Phi_k] = \mathbb{E}[\text{cost}_k(H_k)] + \sum_{i=0}^{k-1} \mathbb{E}[\Phi_{i+1} - \Phi_i \mid \mathcal{E}_i].$$

The first term is at most $8\text{cost}(T^*)$ by Lemma 9. Then, the second term is at most

$$\sum_{i=0}^{k-1} \mathbb{E}[\Phi_{i+1} - \Phi_i \mid \mathcal{E}_i] \leq \sum_{i=0}^{k-1} \frac{\mathbb{E}[\text{cost}_i(H_i)]}{k-i} \leq 8\text{cost}(T^*) \cdot \sum_{i=0}^{k-1} \frac{1}{k-i} \leq 8\text{cost}(T^*) \cdot (1 + \ln k).$$

where we used Lemma 11 and the harmonic sum being at most $(1 + \ln k)$. Putting these two terms together, we are done. finished the proof. □

3.3 Other thoughts

It is a nice exercise to use the same idea to get a *constant factor* approximation by choosing $O(k \log k)$ random centers. This is another variant of k-means, where we allow ourselves more than k centers, but achieve a lower cost clustering.

References

[1] David Arthur and Sergei Vassilvikskii. k-means++: The Advantages of Careful Seeding. SODA 2017