K-MEANS++

Marco Bressan

Università degli Studi di Milano

April 21, 2021
K-means recap

Given a set of $n$ points $X \subset \mathbb{R}^d$, the optimal k-means clustering $C^{OPT}$ is the one given by the set of centroids that minimizes the sum-of-square-residuals $\phi$,

$$
\mathbf{c}^{OPT}_1, \ldots, \mathbf{c}^{OPT}_k = \arg \min_{\mathbf{c}_1, \ldots, \mathbf{c}_k} \phi(\mathbf{c}_1, \ldots, \mathbf{c}_k)
$$

The k-means problem is: given $X$, compute $C^{OPT}$.
Recall: Lloyd’s algorithm has **no approximation guarantee** because of outliers.
K-means recap

Recall: Lloyd’s algorithm has **no approximation guarantee** because of outliers.

Idea: find a better initialisation of centers by **favoring** the outliers.
K-means++


Algorithm 1: K-means++\((X, k)\)

choose a first center, \(c_1\), uniformly at random from \(X\);

for \(i = 2, \ldots, k\) do

\[ \mathbb{P}(c_i = x) = \frac{\min_{j=1,\ldots,i-1} \|x - c_j\|_2^2}{\sum_{x \in X} \min_{j=1,\ldots,i-1} \|x - c_j\|_2^2} \]

end

run Lloyd’s algorithms with initial centers \(c_1, \ldots, c_k\);

return the clustering;
\[ P(c_i = x) = \frac{\min_{j=1,\ldots,i-1} \|x - c_j\|_2^2}{\sum_{x \in X} \min_{j=1,\ldots,i-1} \|x - c_j\|_2^2} \]

You can see that

\[ \min_{j=1,\ldots,i-1} \|x - c_j\|_2^2 \]

is the cost paid by \( x \) in the clustering \( C_{i-1} \) given by the first \( i - 1 \) centers, and

\[ \sum_{x \in X} \min_{j=1,\ldots,i-1} \|x - c_j\|_2^2 \]

is \( \phi(C_{i-1}) \).
Example

\[ p = \frac{1}{n} \]
Example

\[ p = \frac{1}{n} \]

\[ p = \frac{1}{n} \]
Example

\[ p = .45 \quad \text{and} \quad p = .52 \]
Example

\[ p = 0.45 \quad \text{and} \quad p = 0.52 \]
Example

\[ p = 0 \quad \text{and} \quad p = 0.7 \]
Example

$p = 0$

$p = .7$
Example
Example

$X \subset \mathbb{R}^2$, $k = 4$. 

![Scatter plot showing $X \subset \mathbb{R}^2$ with $k = 4$.]
Example

$X \subset \mathbb{R}^2$, $k = 4$. 
Example

$X \subset R^2, \; k = 4.$
Example

$X \subset R^2$, $k = 4$. 

\[
X \subset R^2, \ k = 4. 
\]
$X \subset R^2$, $k = 4$. 
**Theorem.** The clustering $C$ found by K-means++ satisfies:

$$\mathbb{E}[\phi(C)] \leq 8(\ln k + 2) \phi(C_{OPT}).$$

In the remainder we prove a simplified version of the theorem.
We consider the optimal clustering 

$$C^{OPT} = (A_1, \ldots, A_k)$$

and we look at where the centers chosen by k-means++ “land”.
Proof strategy

We consider the optimal clustering

\[ C^{OPT} = (A_1, \ldots, A_k) \]

and we look at where the centers chosen by k-means++ “land”.
We consider the optimal clustering

\[ C^{OPT} = (A_1, \ldots, A_k) \]

and we look at where the centers chosen by k-means++ “land”.

For any cluster \( A \in C^{OPT} \), we denote

\[ \phi_{OPT}(A) = \text{the cost of } A \text{ in } C^{OPT} \]
\[ \phi(A) = \text{the cost of } A \text{ in } C \]
The proof has two parts:

**Part 1:** For any $A \in C^{OPT}$, conditioned on the event that k-means++ chooses a center from $A$, we have:

$$\mathbb{E}[\phi(A)] \leq 8 \phi_{OPT}(A)$$

**Part 2:** In expectation, k-means++ chooses centers from many clusters of $C^{OPT}$. 
Claim 1. For any $A \in C^{OPT}$, conditioned on the event that k-means++ chooses a center from $A$, we have:

$$\mathbb{E}[\phi(A)] \leq 8 \phi_{OPT}(A)$$
Claim 1. For any $A \in C^{OPT}$, conditioned on the event that k-means++ chooses a center from $A$, we have:

$$E[\phi(A)] \leq 8 \phi_{OPT}(A)$$

Proof.

Let $a \in A$ be the random center chosen by k-means++. We consider two cases:

1. $a$ is the first center chosen by k-means++
2. $a$ is not the first center chosen by k-means++
**Case 1:** \(a\) is the first center chosen by k-means++

Then \(a\) is uniform over \(X\). Conditioning on the event \(a \in A\), \(a\) is uniform over \(A\).

\[E[\phi(A)] \leq 8 \phi_{OPT}(A)\]
**Case 1:** \(a\) is the first center chosen by k-means++

Then \(a\) is uniform over \(X\). Conditioning on the event \(a \in A\), \(a\) is uniform over \(A\).

\[
\mathbb{E}[\phi(A)] = \sum_{\hat{a} \in A} \frac{1}{|A|} \cdot \sum_{x \in A} \|x - \hat{a}\|_2^2
\]

\[
\leq 8 \phi_{OPT}(A)
\]
Part 1

Case 1: \( a \) is the first center chosen by \( k \)-means++

Then \( a \) is uniform over \( X \). Conditioning on the event \( a \in A \), \( a \) is uniform over \( A \).

\[
\mathbb{E}[\phi(A)] = \sum_{\hat{a} \in A} \frac{1}{|A|} \cdot \sum_{x \in A} \|x - \hat{a}\|_2^2
\]

\[
= \sum_{\hat{a} \in A} \frac{1}{|A|} \cdot \left( \sum_{x \in A} \|x - \mu\|_2^2 + |A| \cdot \|\hat{a} - \mu\|_2^2 \right)
\]

\[
\leq 8 \phi_{OPT}(A)
\]
Part 1

Case 1: $a$ is the first center chosen by k-means++

Then $a$ is uniform over $X$. Conditioning on the event $a \in A$, $a$ is uniform over $A$.

$$\mathbb{E}[\phi(A)] = \sum_{\hat{a} \in A} \frac{1}{|A|} \cdot \sum_{x \in A} \|x - \hat{a}\|_2^2$$

$$= \sum_{\hat{a} \in A} \frac{1}{|A|} \cdot \left( \sum_{x \in A} \|x - \mu\|_2^2 + |A| \cdot \|\hat{a} - \mu\|_2^2 \right)$$

$$= \sum_{\hat{a} \in A} \frac{1}{|A|} \cdot \sum_{x \in A} \|x - \mu\|_2^2 + \sum_{\hat{a} \in A} \frac{1}{|A|} \cdot |A| \cdot \|\hat{a} - \mu\|_2^2$$

$$\leq 8 \phi_{OPT}(A)$$
Part 1

**Case 1:** $a$ is the first center chosen by k-means++

Then $a$ is uniform over $X$. Conditioning on the event $a \in A$, $a$ is uniform over $A$.

$$\mathbb{E}[\phi(A)] = \sum_{\hat{a} \in A} \frac{1}{|A|} \cdot \sum_{x \in A} \|x - \hat{a}\|_2^2$$

$$= \sum_{\hat{a} \in A} \frac{1}{|A|} \cdot \left( \sum_{x \in A} \|x - \mu\|_2^2 + |A| \cdot \|\hat{a} - \mu\|_2^2 \right)$$

$$= \sum_{\hat{a} \in A} \frac{1}{|A|} \cdot \sum_{x \in A} \|x - \mu\|_2^2 + \sum_{\hat{a} \in A} \frac{1}{|A|} \cdot \|\hat{a} - \mu\|_2^2$$

$$= \sum_{x \in A} \|x - \mu\|_2^2 + \sum_{\hat{a} \in A} \|\hat{a} - \mu\|_2^2$$

$$\leq 8 \phi_{OPT}(A)$$
Part 1

**Case 1:** $a$ is the first center chosen by k-means++

Then $a$ is uniform over $X$. Conditioning on the event $a \in A$, $a$ is uniform over $A$.

$$
\mathbb{E}[\phi(A)] = \sum_{\hat{a} \in A} \frac{1}{|A|} \cdot \sum_{x \in A} \|x - \hat{a}\|^2_2
$$

$$
= \sum_{\hat{a} \in A} \frac{1}{|A|} \cdot \left( \sum_{x \in A} \|x - \mu\|^2_2 + |A| \cdot \|\hat{a} - \mu\|^2_2 \right)
$$

$$
= \sum_{\hat{a} \in A} \frac{1}{|A|} \cdot \sum_{x \in A} \|x - \mu\|^2_2 + \sum_{\hat{a} \in A} \frac{1}{|A|} |A| \cdot \|\hat{a} - \mu\|^2_2
$$

$$
= \sum_{x \in A} \|x - \mu\|^2_2 + \sum_{\hat{a} \in A} \|\hat{a} - \mu\|^2_2
$$

$$
= 2 \sum_{x \in A} \|x - \mu\|^2_2 \leq 8 \phi_{OPT}(A)
$$
Part 1

**Case 1:** \( a \) is the first center chosen by k-means++

Then \( a \) is uniform over \( X \). Conditioning on the event \( a \in A \), \( a \) is uniform over \( A \).

\[
\mathbb{E}[\phi(A)] = \sum_{\hat{a} \in A} \frac{1}{|A|} \cdot \sum_{x \in A} \|x - \hat{a}\|_2^2
\]

\[
= \sum_{\hat{a} \in A} \frac{1}{|A|} \cdot \left( \sum_{x \in A} \|x - \mu\|_2^2 + |A| \cdot \|\hat{a} - \mu\|_2^2 \right)
\]

\[
= \sum_{\hat{a} \in A} \frac{1}{|A|} \cdot \sum_{x \in A} \|x - \mu\|_2^2 + \sum_{\hat{a} \in A} \frac{1}{|A|} |A| \cdot \|\hat{a} - \mu\|_2^2
\]

\[
= \sum_{x \in A} \|x - \mu\|_2^2 + \sum_{\hat{a} \in A} \|\hat{a} - \mu\|_2^2
\]

\[
= 2 \sum_{x \in A} \|x - \mu\|_2^2 = 2 \phi_{OPT}(A) \leq 8 \phi_{OPT}(A)
\]
**Case 2:** \(a\) is not the first center chosen by k-means++

For any \(x \in X\) let \(D(x)^2\) be its squared Euclidean distance from the nearest among the already-chosen centers.
Part 1

**Case 2:** \( a \) is not the first center chosen by k-means++

For any \( x \in X \) let \( D(x)^2 \) be its squared Euclidean distance from the nearest among the already-chosen centers. Conditioning on the event \( a \in A \), we have

\[
P(a = \hat{a}) = \frac{D(\hat{a})^2}{\sum_{x \in A} D(x)^2}
\]
Case 2: \( a \) is not the first center chosen by k-means++

For any \( x \in X \) let \( D(x)^2 \) be its squared Euclidean distance from the nearest among the already-chosen centers. Conditioning on the event \( a \in A \), we have

\[
\mathbb{P}(a = \hat{a}) = \frac{D(\hat{a})^2}{\sum_{x \in A} D(x)^2}
\]

If we choose \( a = \hat{a} \), then the cost of each point \( x \in A \) will be:

\[
\min(D(x)^2, \|x - \hat{a}\|_2^2)
\]
Case 2: \( a \) is not the first center chosen by k-means++

For any \( x \in X \) let \( D(x)^2 \) be its squared Euclidean distance from the nearest among the already-chosen centers. Conditioning on the event \( a \in A \), we have

\[
P(a = \hat{a}) = \frac{D(\hat{a})^2}{\sum_{x \in A} D(x)^2}
\]

If we choose \( a = \hat{a} \), then the cost of each point \( x \in A \) will be:

\[
\min(D(x)^2, \| x - \hat{a} \|_2^2)
\]

Therefore:

\[
\mathbb{E} [\phi(A)] = \sum_{\hat{a} \in A} \frac{D(\hat{a})^2}{\sum_{x \in A} D(x)^2} \sum_{x \in A} \min(D(x)^2, \| x - \hat{a} \|_2^2)
\]
\[ \mathbb{E}[\phi(A)] = \sum_{\hat{a} \in A} \frac{D(\hat{a})^2}{\sum_{x \in A} D(x)^2} \sum_{x \in A} \min(D(x)^2, \|x - \hat{a}\|_2^2) \]
\[ E[\phi(A)] = \sum_{\hat{a} \in A} \frac{D(\hat{a})^2}{\sum_{x \in A} D(x)^2} \sum_{x \in A} \min(D(x)^2, \|x - \hat{a}\|_2^2) \]

Now, for any \( x \in A \), we have the following bound on \( D(\hat{a})^2 \):

\[
D(\hat{a})^2 \leq (D(x) + \|x - \hat{a}\|_2)^2 \quad \text{triangle inequality} \\
\leq 2D(x)^2 + 2\|x - \hat{a}\|_2^2 \quad \text{power-mean ineq: } (b_1 + \ldots + b_m)^2 \leq m(b_1^2 + \ldots + b_m^2)
\]
\[ \mathbb{E}[\phi(A)] = \sum_{\hat{a} \in A} \frac{D(\hat{a})^2}{\sum_{x \in A} D(x)^2} \sum_{x \in A} \min(D(x)^2, \|x - \hat{a}\|_2^2) \sum_{x \in A} D(x)^2 \sum_{x \in A} \min(D(x)^2, \|x - \hat{a}\|_2^2) \]

Now, for any \( x \in A \), we have the following bound on \( D(\hat{a})^2 \):

\[ D(\hat{a})^2 \leq (D(x) + \|x - \hat{a}\|_2)^2 \quad \text{triangle inequality} \]

\[ \leq 2D(x)^2 + 2\|x - \hat{a}\|_2^2 \quad \text{power-mean ineq: } (b_1 + \ldots + b_m)^2 \leq m(b_1^2 + \ldots + b_m^2) \]

By averaging over all \( x \in A \):

\[ D(\hat{a})^2 \leq \frac{1}{|A|} \sum_{x \in A} (2D(x)^2 + 2\|x - \hat{a}\|_2^2) \]
Part 1

\[ \mathbb{E}[\phi(A)] = \sum_{\hat{a} \in A} \frac{D(\hat{a})^2}{\sum_{x \in A} D(x)^2} \sum_{x \in A} \min(D(x)^2, \|x - \hat{a}\|_2^2) \]

Now, for any \( x \in A \), we have the following bound on \( D(\hat{a})^2 \):

\[
D(\hat{a})^2 \leq (D(x) + \|x - \hat{a}\|_2)^2 \quad \text{triangle inequality}
\]

\[
\leq 2D(x)^2 + 2\|x - \hat{a}\|_2^2 \quad \text{power-mean ineq: } (b_1 + \ldots + b_m)^2 \leq m(b_1^2 + \ldots + b_m^2)
\]

By averaging over all \( x \in A \):

\[
D(\hat{a})^2 \leq \frac{1}{|A|} \sum_{x \in A} (2D(x)^2 + 2\|x - \hat{a}\|_2^2)
\]

Thus:

\[
\mathbb{E}[\phi(A)] \leq \sum_{\hat{a} \in A} \frac{\frac{2}{|A|} \sum_{x \in A} (D(x)^2 + \|x - \hat{a}\|_2^2)}{\sum_{x \in A} D(x)^2} \sum_{x \in A} \min(D(x)^2, \|x - \hat{a}\|_2^2)
\]
We’re almost done:

$$
\mathbb{E}[\phi(A)] \leq \sum_{\hat{a} \in A} \frac{2}{|A|} \sum_{x \in A} \frac{(D(x)^2 + \|x - \hat{a}\|_2^2)}{\sum_{x \in A} D(x)^2} \sum_{x \in A} \min(D(x)^2, \|x - \hat{a}\|_2^2)
$$
We’re almost done:

\[
\mathbb{E}[\phi(A)] \leq \sum_{\hat{a} \in A} \frac{2}{|A|} \frac{\sum_{x \in A} (D(x)^2 + \|x - \hat{a}\|^2)}{\sum_{x \in A} D(x)^2} \sum_{x \in A} \min(D(x)^2, \|x - \hat{a}\|^2)
\]

\[
= \frac{2}{|A|} \frac{\sum_{\hat{a} \in A} \sum_{x \in A} D(x)^2}{\sum_{x \in A} D(x)^2} \cdot \sum_{x \in A} \min(D(x)^2, \|x - \hat{a}\|^2)
\]

\[
+ \frac{2}{|A|} \frac{\sum_{\hat{a} \in A} \sum_{x \in A} \|x - \hat{a}\|^2}{\sum_{x \in A} D(x)^2} \cdot \sum_{x \in A} \min(D(x)^2, \|x - \hat{a}\|^2)
\]
We’re almost done:

\[
\mathbb{E}[\phi(A)] \leq \sum_{\hat{a} \in A} \frac{2}{|A|} \frac{\sum_{x \in A} (D(x)^2 + \|x - \hat{a}\|_2^2)}{\sum_{x \in A} D(x)^2} \sum_{x \in A} \min(D(x)^2, \|x - \hat{a}\|_2^2)
\]

\[
= \frac{2}{|A|} \frac{\sum_{\hat{a} \in A} \sum_{x \in A} D(x)^2}{\sum_{x \in A} D(x)^2} \cdot \sum_{x \in A} \min(D(x)^2, \|x - \hat{a}\|_2^2)
\]

\[
+ \frac{2}{|A|} \frac{\sum_{\hat{a} \in A} \sum_{x \in A} \|x - \hat{a}\|_2^2}{\sum_{x \in A} D(x)^2} \cdot \sum_{x \in A} \min(D(x)^2, \|x - \hat{a}\|_2^2)
\]

= 1
We’re almost done:

\[
\mathbb{E}[\phi(A)] \leq \sum_{\hat{a} \in A} \frac{2}{|A|} \frac{\sum_{x \in A} (D(x)^2 + \|x - \hat{a}\|^2)}{\sum_{x \in A} D(x)^2} \sum_{x \in A} \min(D(x)^2, \|x - \hat{a}\|^2)
\]

\[
= \frac{2}{|A|} \frac{\sum_{\hat{a} \in A} \sum_{x \in A} D(x)^2}{\sum_{x \in A} D(x)^2} \cdot \sum_{x \in A} \min(D(x)^2, \|x - \hat{a}\|^2) = 1
\]

\[
+ \frac{2}{|A|} \frac{\sum_{\hat{a} \in A} \sum_{x \in A} \|x - \hat{a}\|^2}{\sum_{x \in A} D(x)^2} \cdot \sum_{x \in A} \min(D(x)^2, \|x - \hat{a}\|^2) \leq 1
\]
We’re almost done:

\[
\mathbb{E}[\phi(A)] \leq \sum_{\hat{a} \in A} \frac{2}{|A|} \frac{\sum_{x \in A} (D(x)^2 + \|x - \hat{a}\|_2^2)}{\sum_{x \in A} D(x)^2} \sum_{x \in A} \min(D(x)^2, \|x - \hat{a}\|_2^2)
\]

\[
= \frac{2}{|A|} \frac{\sum_{\hat{a} \in A} \sum_{x \in A} D(x)^2}{\sum_{x \in A} D(x)^2} \cdot \sum_{x \in A} \min(D(x)^2, \|x - \hat{a}\|_2^2)
\]

\[
+ \frac{2}{|A|} \frac{\sum_{\hat{a} \in A} \sum_{x \in A} \|x - \hat{a}\|_2^2}{\sum_{x \in A} D(x)^2} \cdot \sum_{x \in A} \min(D(x)^2, \|x - \hat{a}\|_2^2)
\]

\[
\leq \frac{4}{|A|} \sum_{\hat{a} \in A} \sum_{x \in A} \|x - \hat{a}\|_2^2
\]
Part 1

We’re almost done:

\[
\mathbb{E}[\phi(A)] \leq \sum_{\hat{a} \in A} \frac{2}{|A|} \sum_{x \in A} \frac{(D(x)^2 + \|x - \hat{a}\|_2^2)}{\sum_{x \in A} D(x)^2} \sum_{x \in A} \min(D(x)^2, \|x - \hat{a}\|_2^2)
\]

\[
= \frac{2}{|A|} \sum_{\hat{a} \in A} \frac{\sum_{x \in A} D(x)^2}{\sum_{x \in A} D(x)^2} \cdot \sum_{x \in A} \min(D(x)^2, \|x - \hat{a}\|_2^2) = 1
\]

\[
+ \frac{2}{|A|} \sum_{\hat{a} \in A} \frac{\sum_{x \in A} \|x - \hat{a}\|_2^2}{\sum_{x \in A} D(x)^2} \cdot \sum_{x \in A} \min(D(x)^2, \|x - \hat{a}\|_2^2) \leq 1
\]

\[
\leq \frac{4}{|A|} \sum_{\hat{a} \in A} \sum_{x \in A} \|x - \hat{a}\|_2^2 \leq 4 \cdot 2\phi_{OPT}(A) = 8\phi_{OPT}(A)
\]
Recap: For any $A \in C^{OPT}$, conditioned on the event that k-means++ chooses a center from $A$, we have:

$$\mathbb{E}[\phi(A)] \leq 8 \phi_{OPT}(A)$$
Part 2

For any $A \in C^{OPT}$, We say that $A$ is **covered** if k-means++ has chosen some center in $A$. Otherwise we say that $A$ is **uncovered**.
For any $A \in C^{OPT}$, We say that $A$ is **covered** if k-means++ has chosen some center in $A$. Otherwise we say that $A$ is **uncovered**.

Thanks to Part 1, we know that covered clusters are “ok” (on them, we pay an almost-optimal cost).
For any $A \in C^{OPT}$, we say that $A$ is **covered** if k-means++ has chosen some center in $A$. Otherwise we say that $A$ is **uncovered**.

Thanks to Part 1, we know that covered clusters are “ok” (on them, we pay an almost-optimal cost).

Therefore we can simplify the model as follows.

**SIMPLIFYING ASSUMPTION**

For all $A \in C^{OPT}$, we have $\phi_{OPT}(A) = 1$.

Moreover, if $A$ is covered then $\phi(A) = \phi_{OPT}(A) = 1$, otherwise $\phi(A) = L \gg 1$. 
Part 2

For any $A \in C^{OPT}$, We say that $A$ is **covered** if k-means++ has chosen some center in $A$. Otherwise we say that $A$ is **uncovered**.

Thanks to Part 1, we know that covered clusters are “ok” (on them, we pay an almost-optimal cost).

Therefore we can simplify the model as follows.

**SIMPLIFYING ASSUMPTION**

For all $A \in C^{OPT}$, we have $\phi_{OPT}(A) = 1$. Moreover, if $A$ is covered then $\phi(A) = \phi_{OPT}(A) = 1$, otherwise $\phi(A) = L \gg 1$.

We will prove: $\mathbb{E}[\phi] \leq \phi_{OPT} \cdot O(\lg k)$
For \( i = 0, \ldots, k \) we denote by \( \phi_i \) the cost of k-means++ after choosing \( i \) centers. By convention \( \mathbb{E}[\phi_0] = \phi_0 = kL \) (think of an initial “external center”).
For $i = 0, \ldots, k$ we denote by $\phi_i$ the cost of k-means++ after choosing $i$ centers. By convention $\mathbb{E}[\phi_0] = \phi_0 = kL$ (think of an initial “external center”).

Now:

$$
\phi_k = \phi_0 + \sum_{i=0}^{k-1} (\phi_{i+1} - \phi_i)
$$
For $i = 0, \ldots, k$ we denote by $\phi_i$ the cost of k-means++ after choosing $i$ centers. By convention $\mathbb{E}[\phi_0] = \phi_0 = kL$ (think of an initial “external center”). Now:

$$\phi_k = \phi_0 + \sum_{i=0}^{k-1} (\phi_{i+1} - \phi_i)$$

Taking expectations:

$$\mathbb{E}[\phi_k] = \mathbb{E}[\phi_0] + \sum_{i=0}^{k-1} (\mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i])$$
Part 2

For $i = 0, \ldots, k$ we denote by $\phi_i$ the cost of k-means++ after choosing $i$ centers. By convention $\mathbb{E}[\phi_0] = \phi_0 = kL$ (think of an initial “external center”).

Now:

$$\phi_k = \phi_0 + \sum_{i=0}^{k-1} (\phi_{i+1} - \phi_i)$$

Taking expectations:

$$\mathbb{E}[\phi_k] = \mathbb{E}[\phi_0] + \sum_{i=0}^{k-1} (\mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i])$$

$$= kL + \sum_{i=0}^{k-1} (\mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i])$$
For \( i = 0, \ldots, k \) we denote by \( \phi_i \) the cost of k-means++ after choosing \( i \) centers. By convention \( \mathbb{E}[\phi_0] = \phi_0 = kL \) (think of an initial “external center”).

Now:

\[
\phi_k = \phi_0 + \sum_{i=0}^{k-1} (\phi_{i+1} - \phi_i)
\]

Taking expectations:

\[
\mathbb{E}[\phi_k] = \mathbb{E}[\phi_0] + \sum_{i=0}^{k-1} (\mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i])
\]

\[
= kL + \sum_{i=0}^{k-1} (\mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i])
\]

\[
= k + \sum_{i=0}^{k-1} ((L - 1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i])
\]
\[
\mathbb{E}[\phi_k] = k + \sum_{i=0}^{k-1} ((L - 1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i])
\]

We can see this as charging round \( i \) with an initial penalty of \( L - 1 \), which the algorithm fights by improving by \( \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i] \).
\[ E[\phi_k] = k + \sum_{i=0}^{k-1} ((L - 1) + E[\phi_{i+1}] - E[\phi_i]) \]

We can see this as charging round \( i \) with an initial penalty of \( L - 1 \), which the algorithm fights by improving by \( E[\phi_{i+1}] - E[\phi_i] \).

Let \( u_i \) the number of uncovered clusters after round \( i \). Note that \( \phi_i = u_i \cdot L + (k - u_i) \).
\[ \mathbb{E}[\phi_k] = k + \sum_{i=0}^{k-1} ((L - 1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i]) \]

We can see this as charging round \( i \) with an initial penalty of \( L - 1 \), which the algorithm fights by improving by \( \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i] \).

Let \( u_i \) the number of uncovered clusters after round \( i \). Note that \( \phi_i = u_i \cdot L + (k - u_i) \).

For any uncovered \( A \), the probability that at round \( i + 1 \) we choose a center from \( A \) is:

\[
\frac{\phi_i(A)}{\phi_i} = \frac{L}{u_i \cdot L + (k - u_i)}
\]
Part 2

\[ E[\phi_k] = k + \sum_{i=0}^{k-1} ((L - 1) + E[\phi_{i+1}] - E[\phi_i]) \]

We can see this as charging round \( i \) with an initial penalty of \( L - 1 \), which the algorithm fights by improving by \( E[\phi_{i+1}] - E[\phi_i] \).

Let \( u_i \) the number of uncovered clusters after round \( i \). Note that \( \phi_i = u_i \cdot L + (k - u_i) \).

For any uncovered \( A \), the probability that at round \( i + 1 \) we choose a center from \( A \) is:

\[ \frac{\phi_i(A)}{\phi_i} = \frac{L}{u_i \cdot L + (k - u_i)} \]

So the probability that we choose a center from some uncovered cluster is:

\[ \frac{u_i \cdot L}{u_i \cdot L + (k - u_i)} \]
The probability that we choose a center from some uncovered cluster is:

$$\frac{u_i \cdot L}{u_i \cdot L + (k - u_i)}$$
The probability that we choose a center from some uncovered cluster is:

\[
\frac{u_i \cdot L}{u_i \cdot L + (k - u_i)} \geq \frac{(k - i) \cdot L}{(k - i) \cdot L + i}
\]
The probability that we choose a center from some uncovered cluster is:

\[
\frac{u_i \cdot L}{u_i \cdot L + (k - u_i)} \geq \frac{(k - i) \cdot L}{(k - i) \cdot L + i}
\]

If this happens (choosing a center from some uncovered cluster), then:

\[
\phi_{i+1} = \phi_i - L + 1 = \phi_i - (L - 1)
\]
The probability that we choose a center from some uncovered cluster is:

\[
\frac{u_i \cdot L}{u_i \cdot L + (k - u_i)} \geq \frac{(k - i) \cdot L}{(k - i) \cdot L + i}
\]

If this happens (choosing a center from some uncovered cluster), then:

\[
\phi_{i+1} = \phi_i - L + 1 = \phi_i - (L - 1)
\]

Therefore:

\[
\mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i] \leq -(L - 1) \cdot \frac{(k - i) \cdot L}{(k - i) \cdot L + i}
\]
Part 2

We’re almost done:

\[
(L - 1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i] \leq (L - 1) - (L - 1) \cdot \frac{(k - i) \cdot L}{(k - i) \cdot L + i}
\]
Part 2

We’re almost done:

\[(L - 1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i] \leq (L - 1) - (L - 1) \cdot \frac{(k - i) \cdot L}{(k - i) \cdot L + i} \]

\[
= (L - 1) \left(1 - \frac{(k - i) \cdot L}{(k - i) \cdot L + i}\right)
\]
We’re almost done:

\[(L - 1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i] \leq (L - 1) - (L - 1) \cdot \frac{(k - i) \cdot L}{(k - i) \cdot L + i}\]

\[= (L - 1) \left( 1 - \frac{(k - i) \cdot L}{(k - i) \cdot L + i} \right)\]

\[= (L - 1) \left( \frac{i}{(k - i) \cdot L + i} \right)\]
Part 2

We’re almost done:

\[(L - 1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i] \leq (L - 1) - (L - 1) \cdot \frac{(k - i) \cdot L}{(k - i) \cdot L + i}\]

\[= (L - 1) \left(1 - \frac{(k - i) \cdot L}{(k - i) \cdot L + i}\right)\]

\[= (L - 1) \left(\frac{i}{(k - i) \cdot L + i}\right)\]

\[< L \cdot \frac{i}{(k - i) \cdot L + i}\]
Part 2

We’re almost done:

\[(L - 1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i] \leq (L - 1) - (L - 1) \cdot \frac{(k - i) \cdot L}{(k - i) \cdot L + i}\]

\[= (L - 1) \left(1 - \frac{(k - i) \cdot L}{(k - i) \cdot L + i}\right)\]

\[= (L - 1) \left(\frac{i}{(k - i) \cdot L + i}\right)\]

\[< \frac{L}{(k - i) \cdot L + i}\]

\[< \frac{k}{(k - i) \cdot L}\]
Part 2

We’re almost done:

\[(L - 1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i] \leq (L - 1) - (L - 1) \cdot \frac{(k - i) \cdot L}{(k - i) \cdot L + i}\]

\[= (L - 1) \left(1 - \frac{(k - i) \cdot L}{(k - i) \cdot L + i}\right)\]

\[= (L - 1) \left(\frac{i}{(k - i) \cdot L + i}\right)\]

\[< \frac{i}{(k - i) \cdot L + i}\]

\[< \frac{k}{(k - i) \cdot L}\]

\[= \frac{k}{k - i}\]
So \((L - 1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i] \leq \frac{k}{k-i}\). Therefore, recalling from before:

\[
\mathbb{E}[\phi_k] = k + \sum_{i=0}^{k-1} \left((L - 1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i]\right)
\]
So \( (L - 1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i] \leq \frac{k}{k-i} \). Therefore, recalling from before:

\[
\mathbb{E}[\phi_k] = k + \sum_{i=0}^{k-1} \left((L - 1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i]\right)
\]

\[
\leq k + \sum_{i=0}^{k-1} \frac{k}{k-i}
\]

This concludes the (simplified) proof that \( \mathbb{E}[\phi] \leq \phi_{\text{OPT}} \cdot O(\ln k) \).
So \((L - 1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i] \leq \frac{k}{k-i}\). Therefore, recalling from before:

\[
\mathbb{E}[\phi_k] = k + \sum_{i=0}^{k-1} ((L - 1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i])
\]

\[
\leq k + \sum_{i=0}^{k-1} \frac{k}{k-i}
\]

\[
= k + k \sum_{i=1}^{k} \frac{1}{i}
\]

21 / 22
So \((L - 1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i] \leq \frac{k}{k-i}\). Therefore, recalling from before:

\[
\mathbb{E}[\phi_k] = k + \sum_{i=0}^{k-1} ((L - 1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i])
\]

\[
\leq k + \sum_{i=0}^{k-1} \frac{k}{k-i}
\]

\[
= k + k \sum_{i=1}^{k} \frac{1}{i}
\]

\[
= k(1 + H_k) \quad H_k \text{ is the } k\text{-th harmonic number}
\]
So \((L - 1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i] \leq \frac{k}{k-i}\). Therefore, recalling from before:

\[
\mathbb{E}[\phi_k] = k + \sum_{i=0}^{k-1} ((L - 1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i])
\]

\[
\leq k + \sum_{i=0}^{k-1} \frac{k}{k-i}
\]

\[
= k + k \sum_{i=1}^{k} \frac{1}{i}
\]

\[
= k(1 + H_k) \quad \text{\(H_k\) is the k-th harmonic number}
\]

\[
\leq k(2 + \ln k)
\]
Part 2

So \((L - 1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i] \leq \frac{k}{k-i}\). Therefore, recalling from before:

\[
\mathbb{E}[\phi_k] = k + \sum_{i=0}^{k-1} ((L - 1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i])
\]

\[
\leq k + \sum_{i=0}^{k-1} \frac{k}{k-i}
\]

\[
= k + k \sum_{i=1}^{k} \frac{1}{i}
\]

\[
= k(1 + H_k) \quad H_k \text{ is the } k\text{-th harmonic number}
\]

\[
\leq k(2 + \ln k)
\]

This concludes the (simplified) proof that \(\mathbb{E}[\phi] \leq \phi_{OPT} \cdot O(\ln k)\).
NOTE!

All the “cleverness” of kmeans++ is in the seeding process: after choosing the centers using the $D^2$ distribution we already have the guarantee $\mathbb{E}[\phi] \leq \phi_{OPT} \cdot O(\ln k)$.

Indeed, we even forgot about running Lloyd’s algorithm after choosing the centers!