K-MEANS++

Marco Bressan

Università degli Studi di Milano

April 21, 2021

K-means recap

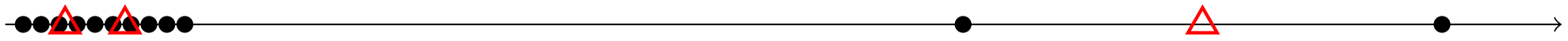
Given a set of n points $X \subset \mathbb{R}^d$, the **optimal k-means clustering** $\mathcal{C}^{OPT}$ is the one given by the set of centroids that minimizes the sum-of-square-residuals ϕ,

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

The k-means problem is: given X, compute $\mathcal{C}^{OPT}$.

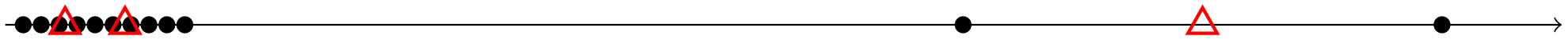
K-means recap

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

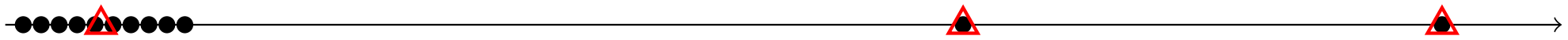


K-means recap

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



Idea: find a better initialisation of centers by **favoring** the outliers.



K-means++

Introduced by Arthur and Vassilvitskii (ACM-SIAM SODA, 2007).

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

choose a first center, $\boldsymbol{c}_1$, uniformly at random from X;

for $i = 2, \ldots, k$ **do**

draw $\boldsymbol{c}_i$ at random from X according to the probability distribution:

$$\mathbb{P}(\boldsymbol{c}_i = \boldsymbol{x}) = \frac{\min_{j=1,\ldots,i-1} \|\boldsymbol{x} - \boldsymbol{c}_j\|_2^2}{\sum_{\boldsymbol{x}\in X} \min_{j=1,\ldots,i-1} \|\boldsymbol{x} - \boldsymbol{c}_j\|_2^2}$$

end

run Lloyd's algorithms with initial centers $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k$;

return the clustering;

K-means++

$$\mathbb{P}(\boldsymbol{c}_i = \boldsymbol{x}) = \frac{\min_{j=1,\dots,i-1} \|\boldsymbol{x} - \boldsymbol{c}_j\|_2^2}{\sum_{\boldsymbol{x}\in X} \min_{j=1,\dots,i-1} \|\boldsymbol{x} - \boldsymbol{c}_j\|_2^2}$$

You can see that

$$\min_{j=1,\dots,i-1} \|\boldsymbol{x} - \boldsymbol{c}_j\|_2^2$$

is the cost paid by $\boldsymbol{x}$ in the clustering $\mathcal{C}_{i-1}$ given by the first $i-1$ centers, and

$$\sum_{\boldsymbol{x}\in X} \min_{j=1,\dots,i-1} \|\boldsymbol{x} - \boldsymbol{c}_j\|_2^2$$

is $\phi(\mathcal{C}_{i-1})$.

Example

Example

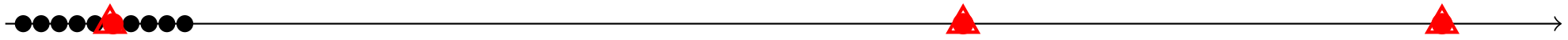
Example

Example

Example

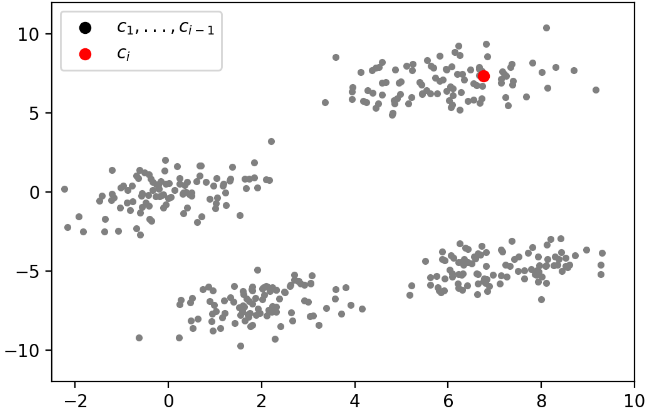
Example

Example



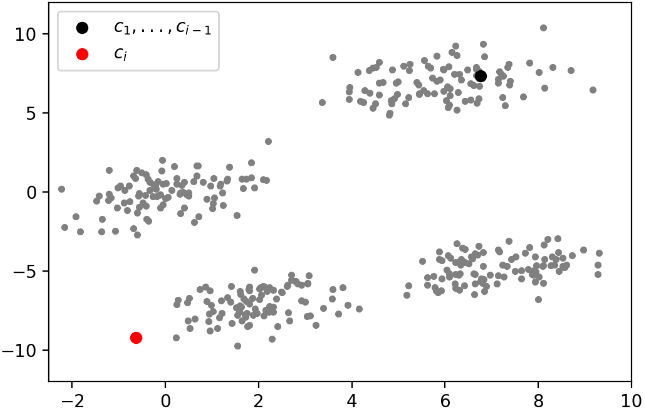
Example

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



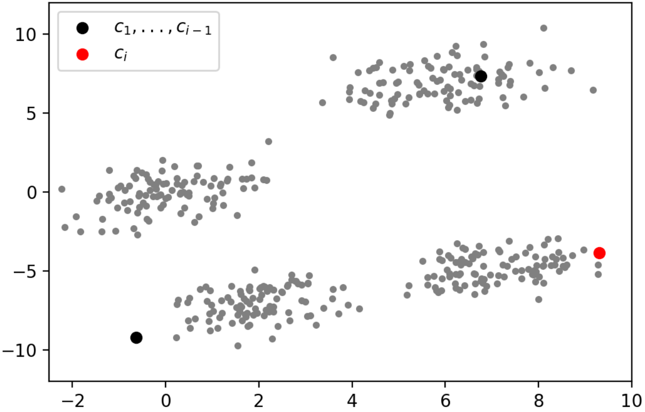
Example

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



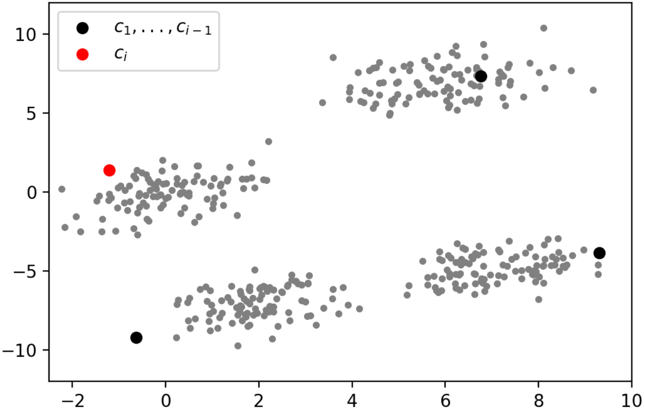
Example

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



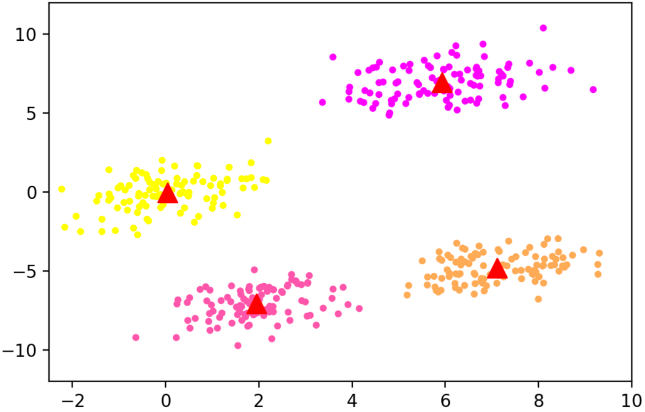
Example

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



Example

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



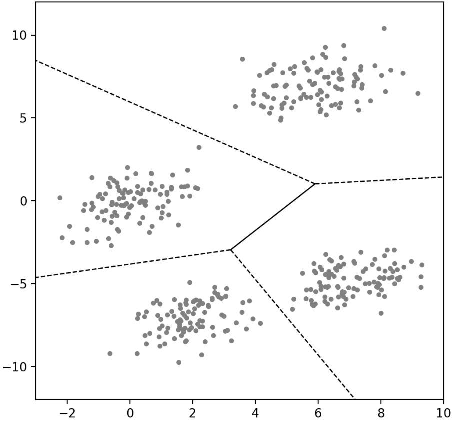
K-means++

Theorem. The clustering $\mathcal{C}$ found by K-means++ satisfies:

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

In the remainder we prove a simplified version of the theorem.

Proof strategy

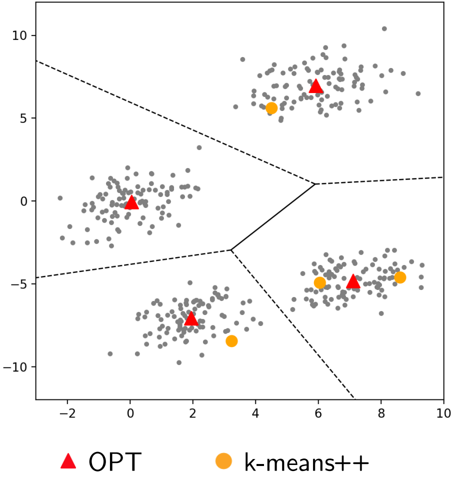


We consider the optimal clustering

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

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

Proof strategy

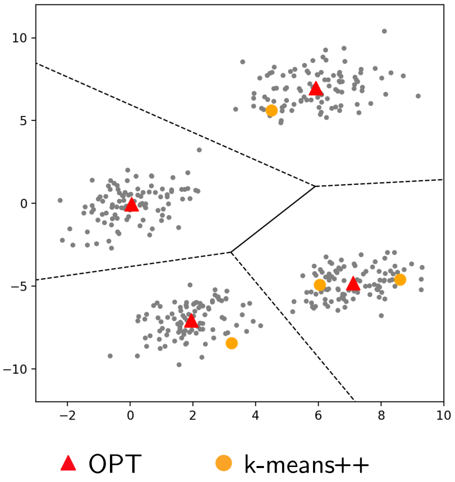


We consider the optimal clustering

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

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

Proof strategy



We consider the optimal clustering

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

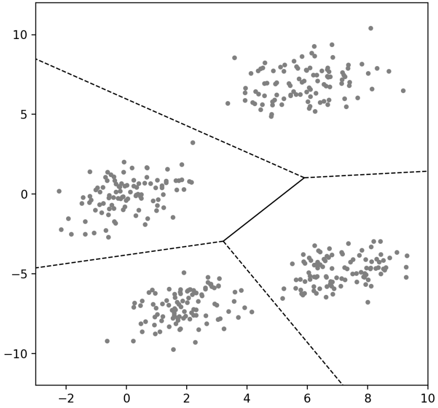
and we look at where the centers chosen by k-means++ “land”.

For any cluster $A \in \mathcal{C}^{OPT}$, we denote

$$\phi_{OPT}(A) = \text{ the cost of } A \text{ in } \mathcal{C}^{OPT}$$

$$\phi(A) = \text{ the cost of } A \text{ in } \mathcal{C}$$

Proof strategy



The proof has two parts:

Part 1: For any $A \in \mathcal{C}^{OPT}$, 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 $\mathcal{C}^{OPT}$.

Part 1

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

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

Part 1

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

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

Proof.

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

1. $\boldsymbol{a}$ is the first center chosen by k-means++
2. $\boldsymbol{a}$ is not the first center chosen by k-means++

Part 1

Case 1: $\boldsymbol{a}$ is the first center chosen by k-means++

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

$$\mathbb{E}[\phi(A)]$$

$$\leq 8\,\phi_{OPT}(A)$$

Part 1

Case 1: $\boldsymbol{a}$ is the first center chosen by k-means++

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

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

$$\leq \ 8\,\phi_{OPT}(A)$$

Part 1

Case 1: $\boldsymbol{a}$ is the first center chosen by k-means++

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

$$\begin{aligned}\mathbb{E}[\phi(A)] &= \sum_{\widehat{\boldsymbol{a}} \in A} \frac{1}{|A|} \cdot \sum_{\boldsymbol{x} \in A} \|\boldsymbol{x} - \widehat{\boldsymbol{a}}\|_2^2 \\ &= \sum_{\widehat{\boldsymbol{a}} \in A} \frac{1}{|A|} \cdot \left(\sum_{\boldsymbol{x} \in A} \|\boldsymbol{x} - \boldsymbol{\mu}\|_2^2 \ + \ |A| \cdot \|\widehat{\boldsymbol{a}} - \boldsymbol{\mu}\|_2^2 \right)\end{aligned}$$

$$\leq \ 8\,\phi_{OPT}(A)$$

Part 1

Case 1: $\boldsymbol{a}$ is the first center chosen by k-means++

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

$$\begin{aligned}
\mathbb{E}[\phi(A)] &= \sum_{\widehat{\boldsymbol{a}} \in A} \frac{1}{|A|} \cdot \sum_{\boldsymbol{x} \in A} \|\boldsymbol{x} - \widehat{\boldsymbol{a}}\|_2^2 \\
&= \sum_{\widehat{\boldsymbol{a}} \in A} \frac{1}{|A|} \cdot \left(\sum_{\boldsymbol{x} \in A} \|\boldsymbol{x} - \boldsymbol{\mu}\|_2^2 \;+\; |A| \cdot \|\widehat{\boldsymbol{a}} - \boldsymbol{\mu}\|_2^2 \right) \\
&= \sum_{\widehat{\boldsymbol{a}} \in A} \frac{1}{|A|} \cdot \sum_{\boldsymbol{x} \in A} \|\boldsymbol{x} - \boldsymbol{\mu}\|_2^2 \;+\; \sum_{\widehat{\boldsymbol{a}} \in A} \frac{1}{|A|} |A| \cdot \|\widehat{\boldsymbol{a}} - \boldsymbol{\mu}\|_2^2
\end{aligned}$$

$$\leq \; 8\,\phi_{OPT}(A)$$

Part 1

Case 1: $\boldsymbol{a}$ is the first center chosen by k-means++

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

$$\begin{aligned}
\mathbb{E}[\phi(A)] &= \sum_{\widehat{\boldsymbol{a}} \in A} \frac{1}{|A|} \cdot \sum_{\boldsymbol{x} \in A} \|\boldsymbol{x} - \widehat{\boldsymbol{a}}\|_2^2 \\
&= \sum_{\widehat{\boldsymbol{a}} \in A} \frac{1}{|A|} \cdot \left(\sum_{\boldsymbol{x} \in A} \|\boldsymbol{x} - \boldsymbol{\mu}\|_2^2 \; + \; |A| \cdot \|\widehat{\boldsymbol{a}} - \boldsymbol{\mu}\|_2^2 \right) \\
&= \sum_{\widehat{\boldsymbol{a}} \in A} \frac{1}{|A|} \cdot \sum_{\boldsymbol{x} \in A} \|\boldsymbol{x} - \boldsymbol{\mu}\|_2^2 \; + \sum_{\widehat{\boldsymbol{a}} \in A} \frac{1}{|A|} |A| \cdot \|\widehat{\boldsymbol{a}} - \boldsymbol{\mu}\|_2^2 \\
&= \sum_{\boldsymbol{x} \in A} \|\boldsymbol{x} - \boldsymbol{\mu}\|_2^2 \; + \sum_{\widehat{\boldsymbol{a}} \in A} \|\widehat{\boldsymbol{a}} - \boldsymbol{\mu}\|_2^2 \\
&\qquad\qquad\qquad \leq \; 8\, \phi_{OPT}(A)
\end{aligned}$$

Part 1

Case 1: $\boldsymbol{a}$ is the first center chosen by k-means++

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

$$
\begin{aligned}
\mathbb{E}[\phi(A)] &= \sum_{\widehat{\boldsymbol{a}} \in A} \frac{1}{|A|} \cdot \sum_{\boldsymbol{x} \in A} \|\boldsymbol{x} - \widehat{\boldsymbol{a}}\|_2^2 \\
&= \sum_{\widehat{\boldsymbol{a}} \in A} \frac{1}{|A|} \cdot \left(\sum_{\boldsymbol{x} \in A} \|\boldsymbol{x} - \boldsymbol{\mu}\|_2^2 \; + \; |A| \cdot \|\widehat{\boldsymbol{a}} - \boldsymbol{\mu}\|_2^2 \right) \\
&= \sum_{\widehat{\boldsymbol{a}} \in A} \frac{1}{|A|} \cdot \sum_{\boldsymbol{x} \in A} \|\boldsymbol{x} - \boldsymbol{\mu}\|_2^2 \; + \; \sum_{\widehat{\boldsymbol{a}} \in A} \frac{1}{|A|} |A| \cdot \|\widehat{\boldsymbol{a}} - \boldsymbol{\mu}\|_2^2 \\
&= \sum_{\boldsymbol{x} \in A} \|\boldsymbol{x} - \boldsymbol{\mu}\|_2^2 \; + \; \sum_{\widehat{\boldsymbol{a}} \in A} \|\widehat{\boldsymbol{a}} - \boldsymbol{\mu}\|_2^2 \\
&= 2 \sum_{\boldsymbol{x} \in A} \|\boldsymbol{x} - \boldsymbol{\mu}\|_2^2 \qquad\qquad \leq \; 8 \, \phi_{OPT}(A)
\end{aligned}
$$

Part 1

Case 1: $\boldsymbol{a}$ is the first center chosen by k-means++

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

$$
\begin{aligned}
\mathbb{E}[\phi(A)] &= \sum_{\widehat{\boldsymbol{a}} \in A} \frac{1}{|A|} \cdot \sum_{\boldsymbol{x} \in A} \|\boldsymbol{x} - \widehat{\boldsymbol{a}}\|_2^2 \\
&= \sum_{\widehat{\boldsymbol{a}} \in A} \frac{1}{|A|} \cdot \left(\sum_{\boldsymbol{x} \in A} \|\boldsymbol{x} - \boldsymbol{\mu}\|_2^2 \, + \, |A| \cdot \|\widehat{\boldsymbol{a}} - \boldsymbol{\mu}\|_2^2 \right) \\
&= \sum_{\widehat{\boldsymbol{a}} \in A} \frac{1}{|A|} \cdot \sum_{\boldsymbol{x} \in A} \|\boldsymbol{x} - \boldsymbol{\mu}\|_2^2 \, + \, \sum_{\widehat{\boldsymbol{a}} \in A} \frac{1}{|A|} |A| \cdot \|\widehat{\boldsymbol{a}} - \boldsymbol{\mu}\|_2^2 \\
&= \sum_{\boldsymbol{x} \in A} \|\boldsymbol{x} - \boldsymbol{\mu}\|_2^2 \, + \, \sum_{\widehat{\boldsymbol{a}} \in A} \|\widehat{\boldsymbol{a}} - \boldsymbol{\mu}\|_2^2 \\
&= 2 \sum_{\boldsymbol{x} \in A} \|\boldsymbol{x} - \boldsymbol{\mu}\|_2^2 \; = \; 2\, \phi_{OPT}(A) \; \leq \; 8\, \phi_{OPT}(A)
\end{aligned}
$$

Part 1

Case 2: $\boldsymbol{a}$ is not the first center chosen by k-means++

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

Part 1

Case 2: $\boldsymbol{a}$ is not the first center chosen by k-means++

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

$$\mathbb{P}(\boldsymbol{a} = \widehat{\boldsymbol{a}}) = \frac{D(\widehat{\boldsymbol{a}})^2}{\sum_{\boldsymbol{x} \in A} D(\boldsymbol{x})^2}$$

Part 1

Case 2: $\boldsymbol{a}$ is not the first center chosen by k-means++

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

$$\mathbb{P}(\boldsymbol{a} = \widehat{\boldsymbol{a}}) = \frac{D(\widehat{\boldsymbol{a}})^2}{\sum_{\boldsymbol{x} \in A} D(\boldsymbol{x})^2}$$

If we choose $\boldsymbol{a} = \widehat{\boldsymbol{a}}$, then the cost of each point $\boldsymbol{x} \in A$ will be:

$$\min(D(\boldsymbol{x})^2, \|\boldsymbol{x} - \widehat{\boldsymbol{a}}\|_2^2)$$

Part 1

Case 2: $\boldsymbol{a}$ is not the first center chosen by k-means++

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

$$\mathbb{P}(\boldsymbol{a} = \widehat{\boldsymbol{a}}) = \frac{D(\widehat{\boldsymbol{a}})^2}{\sum_{\boldsymbol{x} \in A} D(\boldsymbol{x})^2}$$

If we choose $\boldsymbol{a} = \widehat{\boldsymbol{a}}$, then the cost of each point $\boldsymbol{x} \in A$ will be:

$$\min(D(\boldsymbol{x})^2, \|\boldsymbol{x} - \widehat{\boldsymbol{a}}\|_2^2)$$

Therefore:

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

Part 1

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

Part 1

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

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

$$\begin{aligned} D(\widehat{\boldsymbol{a}})^2 &\leq (D(\boldsymbol{x}) + \|\boldsymbol{x} - \widehat{\boldsymbol{a}}\|_2)^2 && \text{triangle inequality} \\ &\leq 2D(\boldsymbol{x})^2 + 2\|\boldsymbol{x} - \widehat{\boldsymbol{a}}\|_2^2 && \text{power-mean ineq: } (b_1 + \ldots + b_m)^2 \leq m(b_1^2 + \ldots + b_m^2) \end{aligned}$$

Part 1

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

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

$$\begin{aligned} D(\widehat{\boldsymbol{a}})^2 &\leq (D(\boldsymbol{x}) + \|\boldsymbol{x} - \widehat{\boldsymbol{a}}\|_2)^2 && \text{triangle inequality} \\ &\leq 2D(\boldsymbol{x})^2 + 2\|\boldsymbol{x} - \widehat{\boldsymbol{a}}\|_2^2 && \text{power-mean ineq: } (b_1 + \ldots + b_m)^2 \leq m(b_1^2 + \ldots + b_m^2) \end{aligned}$$

By averaging over all $\boldsymbol{x} \in A$:

$$D(\widehat{\boldsymbol{a}})^2 \leq \frac{1}{|A|} \sum_{\boldsymbol{x} \in A} \left(2D(\boldsymbol{x})^2 + 2\|\boldsymbol{x} - \widehat{\boldsymbol{a}}\|_2^2\right)$$

Part 1

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

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

$$\begin{aligned} D(\widehat{\boldsymbol{a}})^2 &\leq (D(\boldsymbol{x}) + \|\boldsymbol{x} - \widehat{\boldsymbol{a}}\|_2)^2 && \text{triangle inequality} \\ &\leq 2D(\boldsymbol{x})^2 + 2\|\boldsymbol{x} - \widehat{\boldsymbol{a}}\|_2^2 && \text{power-mean ineq: } (b_1 + \ldots + b_m)^2 \leq m(b_1^2 + \ldots + b_m^2) \end{aligned}$$

By averaging over all $\boldsymbol{x} \in A$:

$$D(\widehat{\boldsymbol{a}})^2 \leq \frac{1}{|A|} \sum_{\boldsymbol{x} \in A} \left(2D(\boldsymbol{x})^2 + 2\|\boldsymbol{x} - \widehat{\boldsymbol{a}}\|_2^2\right)$$

Thus:

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

Part 1

We're almost done:

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

Part 1

We're almost done:

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

$$= \frac{2}{|A|} \frac{\sum_{\widehat{\boldsymbol{a}} \in A} \sum_{\boldsymbol{x} \in A} D(\boldsymbol{x})^2}{\sum_{\boldsymbol{x} \in A} D(\boldsymbol{x})^2} \cdot \sum_{\boldsymbol{x} \in A} \min(D(\boldsymbol{x})^2, \|\boldsymbol{x} - \widehat{\boldsymbol{a}}\|_2^2)$$

$$+ \frac{2}{|A|} \frac{\sum_{\widehat{\boldsymbol{a}} \in A} \sum_{\boldsymbol{x} \in A} \|\boldsymbol{x} - \widehat{\boldsymbol{a}}\|_2^2}{\sum_{\boldsymbol{x} \in A} D(\boldsymbol{x})^2} \cdot \sum_{\boldsymbol{x} \in A} \min(D(\boldsymbol{x})^2, \|\boldsymbol{x} - \widehat{\boldsymbol{a}}\|_2^2)$$

Part 1

We're almost done:

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

$$= \frac{2}{|A|} \frac{\sum_{\widehat{\mathbf{a}} \in A} \sum_{\mathbf{x} \in A} D(\mathbf{x})^2}{\sum_{\mathbf{x} \in A} D(\mathbf{x})^2} \cdot \sum_{\mathbf{x} \in A} \min(D(\mathbf{x})^2, \|\mathbf{x} - \widehat{\mathbf{a}}\|_2^2) \qquad = 1$$

$$+ \frac{2}{|A|} \frac{\sum_{\widehat{\mathbf{a}} \in A} \sum_{\mathbf{x} \in A} \|\mathbf{x} - \widehat{\mathbf{a}}\|_2^2}{\sum_{\mathbf{x} \in A} D(\mathbf{x})^2} \cdot \sum_{\mathbf{x} \in A} \min(D(\mathbf{x})^2, \|\mathbf{x} - \widehat{\mathbf{a}}\|_2^2)$$

Part 1

We're almost done:

$$\mathbb{E}[\phi(A)] \le \sum_{\widehat{\boldsymbol{a}}\in A} \frac{\frac{2}{|A|}\sum_{\boldsymbol{x}\in A}\left(D(\boldsymbol{x})^2 + \|\boldsymbol{x}-\widehat{\boldsymbol{a}}\|_2^2\right)}{\sum_{\boldsymbol{x}\in A} D(\boldsymbol{x})^2} \sum_{\boldsymbol{x}\in A} \min(D(\boldsymbol{x})^2, \|\boldsymbol{x}-\widehat{\boldsymbol{a}}\|_2^2)$$

$$= \frac{2}{|A|} \frac{\sum_{\widehat{\boldsymbol{a}}\in A}\sum_{\boldsymbol{x}\in A} D(\boldsymbol{x})^2}{\sum_{\boldsymbol{x}\in A} D(\boldsymbol{x})^2} \cdot \sum_{\boldsymbol{x}\in A} \min(D(\boldsymbol{x})^2, \|\boldsymbol{x}-\widehat{\boldsymbol{a}}\|_2^2) \qquad = 1$$

$$+ \frac{2}{|A|} \frac{\sum_{\widehat{\boldsymbol{a}}\in A}\sum_{\boldsymbol{x}\in A} \|\boldsymbol{x}-\widehat{\boldsymbol{a}}\|_2^2}{\sum_{\boldsymbol{x}\in A} D(\boldsymbol{x})^2} \cdot \sum_{\boldsymbol{x}\in A} \min(D(\boldsymbol{x})^2, \|\boldsymbol{x}-\widehat{\boldsymbol{a}}\|_2^2) \qquad \le 1$$

Part 1

We're almost done:

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

$$= \frac{2}{|A|} \frac{\sum_{\widehat{\boldsymbol{a}} \in A} \sum_{\boldsymbol{x} \in A} D(\boldsymbol{x})^2}{\sum_{\boldsymbol{x} \in A} D(\boldsymbol{x})^2} \cdot \sum_{\boldsymbol{x} \in A} \min(D(\boldsymbol{x})^2, \|\boldsymbol{x} - \widehat{\boldsymbol{a}}\|_2^2) \qquad = 1$$

$$+ \frac{2}{|A|} \frac{\sum_{\widehat{\boldsymbol{a}} \in A} \sum_{\boldsymbol{x} \in A} \|\boldsymbol{x} - \widehat{\boldsymbol{a}}\|_2^2}{\sum_{\boldsymbol{x} \in A} D(\boldsymbol{x})^2} \cdot \sum_{\boldsymbol{x} \in A} \min(D(\boldsymbol{x})^2, \|\boldsymbol{x} - \widehat{\boldsymbol{a}}\|_2^2) \qquad \leq 1$$

$$\leq \frac{4}{|A|} \sum_{\widehat{\boldsymbol{a}} \in A} \sum_{\boldsymbol{x} \in A} \|\boldsymbol{x} - \widehat{\boldsymbol{a}}\|_2^2$$

Part 1

We're almost done:

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

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

$$+ \frac{2}{|A|}\frac{\sum_{\hat{\boldsymbol{a}}\in A}\sum_{\boldsymbol{x}\in A} \|\boldsymbol{x}-\hat{\boldsymbol{a}}\|_2^2}{\sum_{\boldsymbol{x}\in A} D(\boldsymbol{x})^2} \cdot \sum_{\boldsymbol{x}\in A} \min(D(\boldsymbol{x})^2, \|\boldsymbol{x}-\hat{\boldsymbol{a}}\|_2^2) \qquad \le 1$$

$$\le \frac{4}{|A|}\sum_{\hat{\boldsymbol{a}}\in A}\sum_{\boldsymbol{x}\in A} \|\boldsymbol{x}-\hat{\boldsymbol{a}}\|_2^2 \;\le\; 4\cdot 2\phi_{OPT}(A) \;=\; 8\phi_{OPT}(A)$$

Part 1

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

Part 2

For any $A \in \mathcal{C}^{OPT}$, 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).

Part 2

For any $A \in \mathcal{C}^{OPT}$, 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 $A \in \mathcal{C}_{OPT}$, 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 $A \in \mathcal{C}^{OPT}$, 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 $A \in \mathcal{C}_{OPT}$, 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: $\mathbb{E}[\phi] \leq \phi_{OPT} \cdot O(\lg k)$

Part 2

For $i = 0, \ldots, k$ we denote by ϕ_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").

Part 2

For $i = 0, \ldots, k$ we denote by ϕ_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:

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

Part 2

For $i = 0, \ldots, k$ we denote by ϕ_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 ϕ_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:

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

Part 2

For $i = 0, \ldots, k$ we denote by ϕ_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:

$$\begin{aligned}
\mathbb{E}[\phi_k] &= \mathbb{E}[\phi_0] + \sum_{i=0}^{k-1} (\mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i]) \\
&= kL + \sum_{i=0}^{k-1} (\mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i]) \\
&= k + \sum_{i=0}^{k-1} ((L-1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i])
\end{aligned}$$

Part 2

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

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

Part 2

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

We can see this as charging round i with an initialy 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 $\phi_i = u_i \cdot L + (k - u_i)$.

Part 2

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

We can see this as charging round i with an initialy 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

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

We can see this as charging round i with an initialy 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:

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

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

$$\frac{u_i \cdot L}{u_i \cdot L + (k - u_i)}$$

Part 2

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

$$\frac{u_i \cdot L}{u_i \cdot L + (k - u_i)}$$

Part 2

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

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

Part 2

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:

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

Part 2

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:

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

Part 2

We're almost done:

$$\begin{aligned}(L-1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i] &\le (L-1) - (L-1)\cdot\frac{(k-i)\cdot L}{(k-i)\cdot L + i} \\ &= (L-1)\left(1 - \frac{(k-i)\cdot L}{(k-i)\cdot L + i}\right)\end{aligned}$$

Part 2

We're almost done:

$$
\begin{aligned}
(L-1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i] &\leq (L-1) - (L-1) \cdot \frac{(k-i) \cdot L}{(k-i) \cdot L + i} \\
&= (L-1) \left(1 - \frac{(k-i) \cdot L}{(k-i) \cdot L + i} \right) \\
&= (L-1) \left(\frac{i}{(k-i) \cdot L + i} \right)
\end{aligned}
$$

Part 2

We're almost done:

$$\begin{aligned}
(L-1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i] &\leq (L-1) - (L-1) \cdot \frac{(k-i) \cdot L}{(k-i) \cdot L + i} \\
&= (L-1)\left(1 - \frac{(k-i) \cdot L}{(k-i) \cdot L + i}\right) \\
&= (L-1)\left(\frac{i}{(k-i) \cdot L + i}\right) \\
&< L \frac{i}{(k-i) \cdot L + i}
\end{aligned}$$

Part 2

We're almost done:

$$
\begin{aligned}
(L-1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i] &\leq (L-1) - (L-1) \cdot \frac{(k-i) \cdot L}{(k-i) \cdot L + i} \\
&= (L-1)\left(1 - \frac{(k-i) \cdot L}{(k-i) \cdot L + i}\right) \\
&= (L-1)\left(\frac{i}{(k-i) \cdot L + i}\right) \\
&< L\frac{i}{(k-i) \cdot L + i} \\
&< L\frac{k}{(k-i) \cdot L}
\end{aligned}
$$

Part 2

We're almost done:

$$
\begin{aligned}
(L-1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i] &\leq (L-1) - (L-1) \cdot \frac{(k-i)\cdot L}{(k-i)\cdot L + i} \\
&= (L-1)\left(1 - \frac{(k-i)\cdot L}{(k-i)\cdot L + i}\right) \\
&= (L-1)\left(\frac{i}{(k-i)\cdot L + i}\right) \\
&< L\frac{i}{(k-i)\cdot L + i} \\
&< L\frac{k}{(k-i)\cdot L} \\
&= \frac{k}{k-i}
\end{aligned}
$$

Part 2

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

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

Part 2

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

$$\begin{aligned}\mathbb{E}[\phi_k] &= k+\sum_{i=0}^{k-1}\left((L-1)+\mathbb{E}[\phi_{i+1}]-\mathbb{E}[\phi_i]\right)\\ &\le k+\sum_{i=0}^{k-1}\frac{k}{k-i}\end{aligned}$$

Part 2

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

$$\begin{aligned}
\mathbb{E}[\phi_k] &= k + \sum_{i=0}^{k-1} \left((L-1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i]\right) \\
&\leq k + \sum_{i=0}^{k-1} \frac{k}{k-i} \\
&= k + k \sum_{i=1}^{k} \frac{1}{i}
\end{aligned}$$

Part 2

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

$$\begin{aligned}
\mathbb{E}[\phi_k] &= k + \sum_{i=0}^{k-1} \left((L-1)+\mathbb{E}[\phi_{i+1}]-\mathbb{E}[\phi_i]\right) \\
&\leq k + \sum_{i=0}^{k-1} \frac{k}{k-i} \\
&= k + k\sum_{i=1}^{k} \frac{1}{i} \\
&= k(1+H_k) \qquad H_k \text{ is the } k\text{-th harmonic number}
\end{aligned}$$

Part 2

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

$$\begin{aligned}
\mathbb{E}[\phi_k] &= k + \sum_{i=0}^{k-1} \left((L-1) + \mathbb{E}[\phi_{i+1}] - \mathbb{E}[\phi_i]\right) \\
&\leq k + \sum_{i=0}^{k-1} \frac{k}{k-i} \\
&= k + k \sum_{i=1}^{k} \frac{1}{i} \\
&= k(1 + H_k) && H_k \text{ is the } k\text{-th harmonic number} \\
&\leq k(2 + \ln k)
\end{aligned}$$

Part 2

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

$$
\begin{aligned}
\mathbb{E}[\phi_k] &= k + \sum_{i=0}^{k-1}\left((L-1)+\mathbb{E}[\phi_{i+1}]-\mathbb{E}[\phi_i]\right) \\
&\le k + \sum_{i=0}^{k-1}\frac{k}{k-i} \\
&= k + k\sum_{i=1}^{k}\frac{1}{i} \\
&= k(1+H_k) \qquad\qquad H_k \text{ is the } k\text{-th harmonic number} \\
&\le k(2+\ln k)
\end{aligned}
$$

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

K-means++

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!