Instead of arbitrarily initializing cluster centers in Lloyd’s k-means algorithm, k-mean++ algorithm chooses a center using a probabilistic version of farthest-first traversal.

Second part of the lecture covers impossibility theorem which states that no clustering function satisfies all 3 axioms.

\textbf{\textit{k-means++}}

Here’s Lloyd’s k-means algorithm:

\begin{algorithm}
\begin{algorithmic}
  \Require $x_1, \ldots, x_n \in \mathbb{R}^d$, $k \in \mathbb{N}$
  \State Arbitrarily initialize $k$ centers $c_1, c_2, \ldots, c_k \in \mathbb{R}^d$
  \State Assign each $x_i$ to the closest $C_j$ (this creates a partition $P_1, P_2, \ldots, P_k$)
  \State Re-compute centers $c_j = \frac{1}{|P_j|} \sum_{x_i \in P_j} x_i$
  \State Repeat step 2-3 until convergence (up to some $\epsilon$)
\end{algorithmic}
\end{algorithm}

\textbf{Fact 1.} \textit{Solution to Lloyd’s method for k-means can be arbitrarily worse from the optimal solution.}

\textbf{Lemma 2.} \textit{At every step of the algorithm, the k-means cost can only improve.}

\textit{Proof.} For step 2, observe that the $x_i$ are assigned to their closest centers; if a point is assigned to another center, cost would go up. For step 3, once the partition is fixed, the centroid $\frac{1}{|P_j|} \sum_{x_i \in P_j} x_i$ minimizes cost.

\textbf{Possible Initialization Methods}

- Choose centers uniformly at random: Suppose we were given data that is evenly distributed across $k$ natural clusters ($P_1, \ldots, P_k$). If the initial centers ($c_1, \ldots, c_k$) are chosen uniformly at random from data points $x_1, \ldots, x_n$, then the probability that we choose an initial center from each cluster is low.

  This is the \textit{coupon collector problem}. The average time of number of draws to select at least one data point from each cluster is $k \log k$. We could modify the algorithm to run $(k \log k)$-means then merge clusters down to $k$, but this increases time complexity.

  \textit{Lower bound of cost:} Consider an optimal clustering situation, where the $n$ data points are even distributed across $k$ clusters in $\mathbb{R}^1$. The clusters are of radius $\delta$ while the clusters are at least a distance $B$ away from any other. $B$ can be chosen to be much larger than $\delta$. 
On initialization, we select \( k \) points from \( x_1, \ldots, x_n \) uniformly at random. It turns out the expected number of draws to select a one data point from each of the \( k \) clusters is highly concentrated around \( k \log k \). Therefore, with high probability, there will be a cluster that is not represented in the initialization centers. After running \( k \)-means, it is likely that one of the centroids straddles two clusters, and so is on average around a distance of \( B/2 \) from the data points it represents. It follows that the cost of the output is \( \Omega(B^2n) \), where on the other hand, the optimal cost is \( O(\delta^2n) \). And so, uniform random initialization can have arbitrarily worse cost than optimal.

- Farthest-First Traversal: If there are outliers in the data, then this method can have arbitrarily worse cost than optimal. Exercise: why?

- Probabilistic Farthest-First Traversal (\( k \)-means++ paper):

\begin{algorithm}
\textbf{Algorithm 2} \( k \)-means++ initialization algorithm
\begin{enumerate}
    \item Uniformly at random pick \( C_1 \) from \( x_1, \ldots, x_n \).
    \item Let \( C = \{c_1\} \).
    \item Assign \( x_j \) probability \( p_j := \frac{1}{Z}d(x_j, C)^2 \), where \( d(x, C) \) is the usual distance from a point to a set and \( Z \) is an appropriate normalization factor.
    \item Select a point \( x \) according to the probabilities \( p_j \) and let \( C \leftarrow C \cup \{c\} \).
    \item Repeat step 3-4 until \( |C| = k \).
\end{enumerate}
\end{algorithm}

\textbf{Theorem 3.} The \( k \)-means++ algorithm, using the above initialization, obtains expected cost:

\[ \mathbb{E}[\text{cost}(c)] \leq O(\log k) \cdot \text{OPT}, \]

where OPT is the cost of an optimal clustering.

In the following, let \( A \subset X = \{x_1, \ldots, x_n\} \) be a subset, \( C = \{c_1, \ldots, c_k\} \) be the cluster centroids. We define the following notation:

\[ \phi_C(A) = \sum_{a \in A} \min_{c_j \in C} ||a - c_j||^2 \quad \phi = \phi_C(X) = \text{cost}(C) \]

\[ \phi_{\text{opt}}(A) = \sum_{a \in A} \min_{c_j \in C_{\text{opt}}} ||a - c_j||^2 \quad \phi_{\text{opt}} = \phi_{\text{opt}}(X) \]

\textit{Conceptual proof}.
agenda 1. if the first pick falls under region 2, what expected cost for region 2 would be?

agenda 2. if some points are already picked, what expected cost for a particular region would be for next pick?

**Lemma 4** (Lemma 3.2 in $k$-means++ paper). Let $A$ be a cluster from $C_{opt}$, let $C$ be just one cluster chosen uniformly at random from $A$. Then $E[\phi(A)] \leq 2\phi_{opt}(A)$.

**Proof.**

\[
E[\phi(A)] = \frac{1}{|A|} \sum_{a_0 \in A} \sum_{a \in A} ||a - a_0||^2, \text{ } a_0 \text{ is a center that chosen uniformly at random from } A
\]

\[
= \frac{1}{|A|} \sum_{a_0 \in A, a \in A} ||a - a_0||^2, \text{ (recall that } E[||x - y||^2] = 2E[||x - E(x)||^2])
\]

\[
= 2 \sum_{a \in A} ||a - \frac{1}{|A|} \sum_{a \in A} a||^2
\]

\[
= 2 \sum_{a \in A} ||a - c(A)||^2
\]

\[
\leq 2\phi_{opt}(A)
\]

**Lemma 5** (Lemma 3.3 in $k$-means++ paper). Let $A$ be an arbitrary cluster from $C_{opt}$ and $C$ be some arbitrary clustering. If we add a random center to $C$ ($C$ is a set of centers) from $A$ according to $k$-means++ weighting, then $E[\phi(A)] \leq 8\phi_{opt}(A)$.

**Proof.**

Observation: probability that $a_0 \in A$ is chosen: $D^2(a_0)/\sum_{a \in A} D^2(a)$, where $D^2(a_0) = d^2(a_0, C)$ and $D(a_0)$ denotes the shortest distance from $a_0$ to the closest center we have already chosen.

For a given point $a \in A$, after choosing the center $a_0$, the contribution of $a$ to the cost will be
min(D^2(a), ||a - a_0||^2).

\[ \mathbb{E} [\phi(A)] = \sum_{a_0 \in A} \frac{D^2(a_0)}{\sum_{a \in A} D^2(a)} \sum_{a \in A} \min(D^2(a), ||a - a_0||^2) \]

\[ D(a_0) \leq D(a) + ||a - a_0||, \forall a, a_0 \]
\[ D^2(a_0) \leq (D(a) + ||a - a_0||)^2 \]
\[ \leq 2D^2(a) + 2||a - a_0||^2 \]
\[ \sum_{a \in A} D^2(a_0) \leq 2 \sum_{a \in A} (D^2(a) + ||a - a_0||^2) \]
\[ D^2(a_0) \leq \frac{2}{|A|} \sum_{a \in A} D^2(a) + \frac{2}{|A|} \sum_{a \in A} ||a - a_0||^2 \]

\[ \mathbb{E} [\phi(A)] \leq \frac{2}{|A|} \sum_{a_0 \in A} \sum_{a} \frac{D^2(a)}{\sum_{a} D^2(a)} \sum_{a \in A} \min(D^2(a), ||a - a_0||^2) + \]
\[ \frac{2}{|A|} \sum_{a_0 \in A} \sum_{a} \frac{||a - a_0||^2}{\sum_{a} D^2(a)} \sum_{a \in A} \min(D^2(a), ||a - a_0||^2) \]

(pick ||a - a_0||^2 for the first term and pick D^2(a) for the second term.)
\[ \leq \frac{4}{|A|} \sum_{a_0 \in A} \sum_{a} ||a - a_0||^2 \]
\[ \leq 4 \cdot 2\phi_{opt}(A) = 8\phi_{opt}(A) \]

Lemma 6 (Lemma 3.4 in k-means++ paper). Let C be any arbitrary clustering we have chosen, choose \( u > 0 \) (number of uncovered clustering from \( C_{opt} \)). The corresponding uncovered points are \( \chi_u \). Let \( \chi_c = \chi - \chi_u \).

\[ \text{Figure 2: Optimal Clustering with } k = 5 \]
Now suppose we add $t \leq u$ random centers (according to $k$-means++) and $C' = C \cup \{c_1, c_2, \ldots, c_t\}$. The corresponding cost is $\phi'$.

\[
\mathbb{E}[\phi'] \leq (\phi(x_c) + 8\phi_{opt}(x_u))(1 + H_t) + \frac{u-t}{u} \phi(x_u)
\]

$H_t = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{t}$ is the harmonic sum

Let $A$ be the cluster which is covered by the first pick. Then $u = k - 1$, chose $t = k - 1$

\[
\mathbb{E}[\phi'] \leq (\phi(A) + 8\phi_{opt}(\chi - A)(1 + H_{k-1}), \text{ (where } H_{k-1} = 2 + \log k) \\
= (\phi(A) + 8\phi_{opt} - 8\phi_{opt}(A))(2 + \log k), \text{ (where } \phi(A) \leq 2\phi_{opt}(A)) \\
\leq (2 + \log k)8\phi_{opt}
\]

Proof. The proof was done by induction, showing that if we can prove equations $P(u, t - 1)$ and $P(u - 1, t - 1)$ hold true, then $P(u, t)$ holds true. Base case of the induction is $P(u, 0)$ for $u > 0$ and $P(1, 1)$.

Example: Let $k = 3$ for optimal clustering, after first pick there are 2 uncovered clusters from $C_{opt}$. Now we want to prove that $P(2, 2)$ holds true when we pick other two centers with $D^2$ weighting. By induction, if we prove that $P(2, 1)$ and $P(1, 1)$ hold true then the result holds.

Algorithm 3 Local Swap Algorithm

\begin{algorithm}
\textbf{input} $x_1 \cdots x_n$, $k \in \mathbb{N}$
1: Pick $T \subset \{x_1 \cdots x_n\}$, $|T| = k$
2: Swap $x_i \in T$ with $x_j \in X$ if it improves the k-means cost.
3: Repeat step 2 until no more improvement can be made.
\end{algorithm}

Lemma 7 (w/o proof). The solution to "local swap method" is no more than 25 optimum.

\[k\text{-means++ & Lloydss } \sim O(\log k) \cdot OPT\]

\[\text{Local Swap Method } \sim 25 \cdot OPT\]

Kleinberg’s Impossibility Theorem for Clustering

Here, we take an axiomatic approach to clustering: define a clustering procedure as an algorithm $f(X, d)$ that takes in data $X$ and metric on $X$, $d$, and returns a partition of $X$. That is, $f(X, d) = \{X_1, \ldots, X_k\}$, where $X = X_1 \sqcup \cdots \sqcup X_k$. The following are natural qualities we might hope our clustering algorithm to satisfy:

1. Scale-Invariance

Choice of unit should not affect clustering

\[f(x, d) = f(x, \alpha \cdot d), \text{ for any } \alpha > 0\]
2. Richness

Different d’s can give different partitions. In fact, for all partitions $\mathcal{P}$, there should exist some metric $d$ yielding that partition, $f(X, d) = \mathcal{P}$.

3. Consistency

If $d$ yields a partition $\mathcal{P}$, then if $\bar{d}$ is a metric that only reduces distances within clusters and increases distances between clusters, then $f(X, d) = f(X, \bar{d})$. That is, if $f(X, d) = \mathcal{P}$, and

\[
\begin{align*}
\bar{d}(i, j) &\leq d(i, j) \text{ for } i, j \text{ in the same cluster} \\
\bar{d}(i, j) &\geq d(i, j) \text{ for } i, j \text{ in different clusters},
\end{align*}
\]

then $f(X, d) = f(X, \bar{d})$.

**Theorem 8.** There exists no $f$ which satisfies axiom 1, 2 & 3.

**Proof.** Suppose there is a set of three points $\{x_1, x_2, x_3\}$. Two distance function $d$ and $d'$ such that $f(x, d)$ gives a clustering of $\{\{x_1\}, \{x_2\}, \{x_3\}\}$ and $f(x, d')$ gives a clustering of $\{\{x_1, x_2\}, \{x_3\}\}$.

It can be observed that

\[f(x, d) \neq f(x, d')\]  \hspace{1cm} (1)

By scale-invariance,

\[f(x, \alpha \cdot d') = f(x, d')\]  \hspace{1cm} (2)

We can find an $\alpha$ that $\alpha \cdot d'$ enlarges distance between any two points. If consistency holds, it means new distance function $\alpha \cdot d'$ shouldn’t change partition result of $f(x, d)$ because $\alpha \cdot d'$ increases all between-cluster distances. However, from (1) and (2) we know that $f(x, d) \neq f(x, \alpha \cdot d')$, so consistency doesn’t hold for the partition function $f$.

Note that the $k$-means algorithm is not rich because it can only yield $k$ clusters.

**References**
