COMPSCI 690RA: Randomized Algorithms and Probabilistic Data Analysis

Prof. Cameron Musco

University of Massachusetts Amherst. Spring 2022.
Lecture 6

Logistics (Lots of Them)

$\boxed{2.3}$

- Problem Set 2 is due tomorrow 3/3 at 8pm.
- One page project proposal due Monday 3/7.
- Midterm next week in class – designed to be 1.5 hours long, but I will give the full class for it.
- Closed book, mostly short-answer style questions.
- See Schedule tab for midterm study guide/practice questions.
- I will hold additional office hours Monday 3/7 from 4-6pm for midterm review.
- We again do not have a quiz this week due to the upcoming midterm.

Summary

Last Time:

$O(n \log n)$

- Saw how ℓ_0 sampling can be used to solve connectivity using $O(n \log^c n)$ bits of memory in a streaming setting.

AB

- Approximate matrix multiplication via non-unifom norm-based sampling. Analysis via outer-product view of matrix multiplication + linearity of variance.

Ax

A

- Stochastic trace estimation – Hutchinson's method and its full analysis via linearity of variance for pairwise-independent random variables. $x^T A x$

Summary

Last Time:

- Saw how ℓ_0 sampling can be used to solve connectivity using $O(n \log^c n)$ bits of memory in a streaming setting.
- Approximate matrix multiplication via non-unifom norm-based sampling. Analysis via outer-product view of matrix multiplication + linearity of variance.
- Stochastic trace estimation – Hutchinson's method and its full analysis via linearity of variance for pairwise-independent random variables.

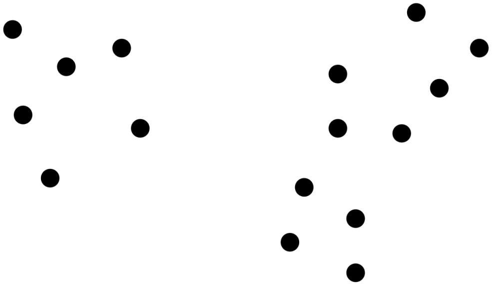
Today: More applications of non-uniform and adaptive sampling to clustering and low-rank approximation.

- The k-means++ algorithm and its analysis.
- Randomized low-rank approximation via norm-based sampling, building on approximate matrix multiplication analysis.

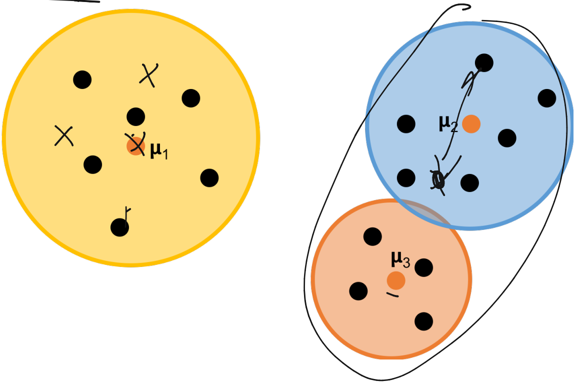
k-means clustering and k-means ++

k-means Clustering

Given $x_1, \ldots, x_n \in \mathbb{R}^d$, assign to clusters $\{C_1, \ldots, C_k\}$ to minimize $\sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|_2^2$ where $\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$ is the cluster centroid.

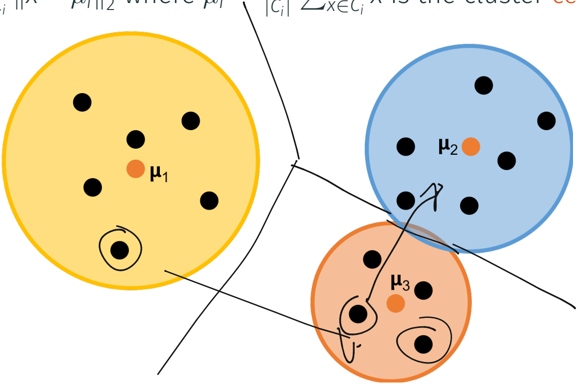


Given $x_1, \ldots, x_n \in \mathbb{R}^d$, assign to clusters $\{C_1, \ldots, C_k\}$ to minimize $\sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|_2^2$ where $\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$ is the cluster centroid.



k-means Clustering

Given $x_1, \ldots, x_n \in \mathbb{R}^d$, assign to clusters $\{C_1, \ldots, C_k\}$ to minimize $\sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|_2^2$ where $\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$ is the cluster centroid.



Probably the most popular clustering objective in practice. But minimizing it is surprisingly hard! $O(n^{dk+1})$ time is the best known for exact minimization, and assuming $P \neq NP$, the exponential dependences on k, d are necessary.

Lloyd's Algorithm

In practice k-means clustering is almost always solved with alternating minimization.

Lloyd's Algorithm

In practice k-means clustering is almost always solved with alternating minimization.

Lloyd's Algorithm: k is an input to algo.

1. Initialize some set of clusters $\{C_1, \ldots, C_k\}$ with centroids $\mu_1, \ldots, \mu_k$.
2. Reassign each datapoint x_i to cluster C_j where $j = \arg\min_{j \in [k]} \|x_i - \mu_j\|_2^2$.
3. Recompute centroids $\mu_1, \ldots, \mu_k$ to reflect the new clusters.
4. Repeat (2)-(3).

Lloyd's Algorithm

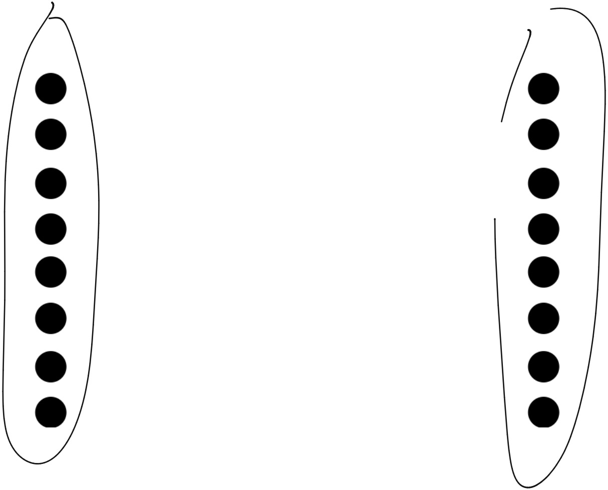
In practice k-means clustering is almost always solved with alternating minimization.

Lloyd's Algorithm:

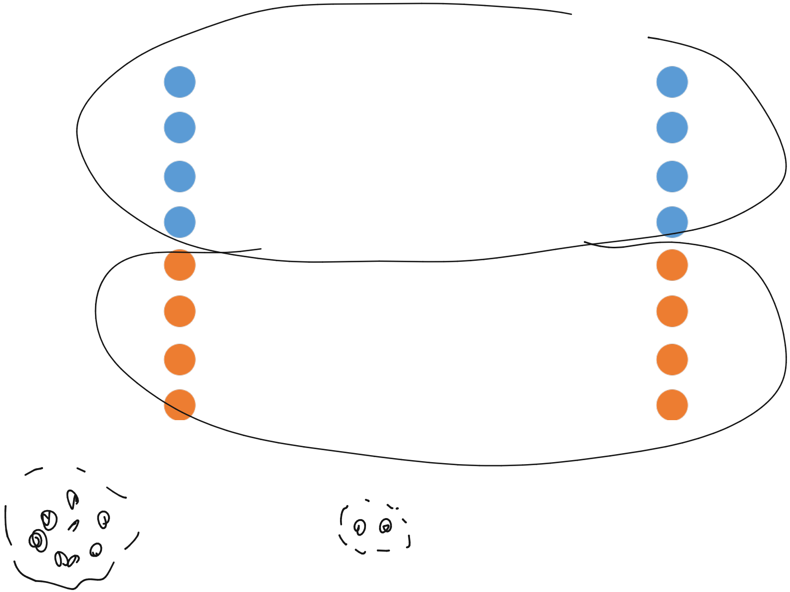
1. Initialize some set of clusters $\{C_1, \ldots, C_k\}$ with centroids $\mu_1, \ldots, \mu_k$.
2. Reassign each datapoint x_i to cluster C_j where $j = \arg\min_{j \in [k]} \|x_i - \mu_j\|_2^2$.
3. Recompute centroids $\mu_1, \ldots, \mu_k$ to reflect the new clusters.
4. Repeat (2)-(3).

Observe that the cost of the clustering can never increase. However, if the initialization is bad, can get caught in a bad local minimum.

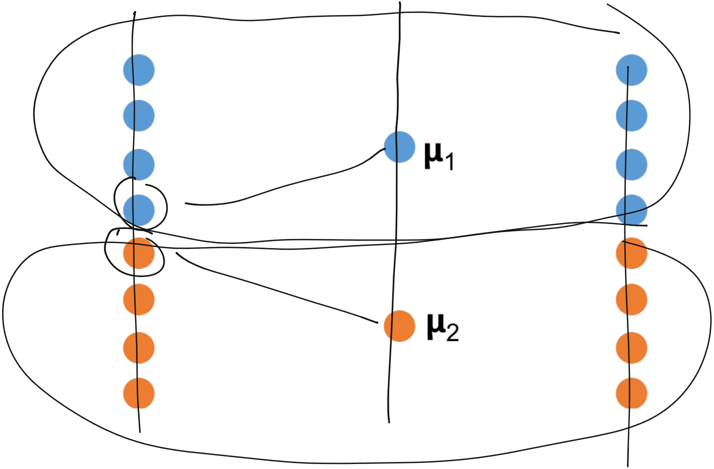
Lloyd's Algorithm



Lloyd's Algorithm



Lloyd's Algorithm



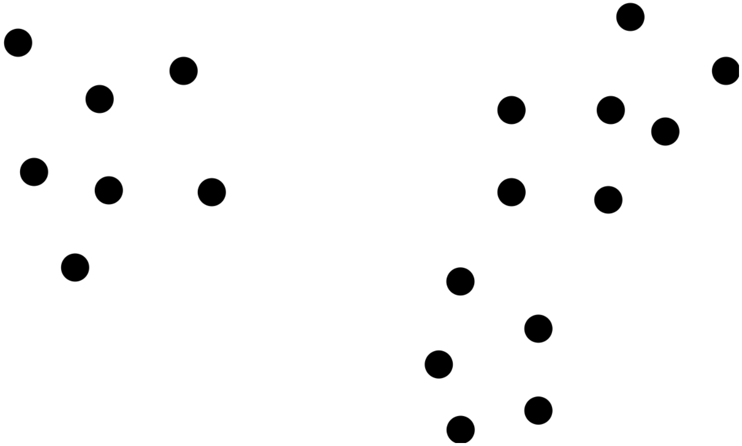
k-means++

k-means++: An extremely simple randomized initialization scheme for k-means which yields a $O(\log k)$ approximation to the optimal clustering.

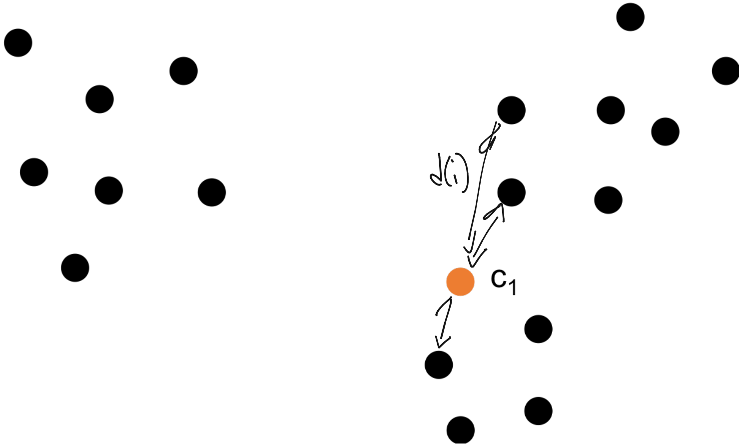
k-means++: An extremely simple randomized initialization scheme for k-means which yields a $O(\log k)$ approximation to the optimal clustering. $x_1 \ldots x_n$

- Initialize probabilities $p_i = 1/n$ for $i \in [n]$.
- Initialize list of cluster centers $C = \{\}$.
- For $j = 1, 2, \ldots k$
  - Set center $c_j \in \{x_1, \ldots, x_n\}$ to x_i with probability p_i. Add c_j to C.
  - For all $i \in [n]$, let $d(i) = \min_{c \in C} \|x_i - c\|_2^2$.
  - For all $i \in [n]$, let $p_i = d(i) / \sum_{i=1}^n d(i)$.
- Let $C_1, \ldots, C_k$ be the clusters formed by assigning each data point to the nearest center in $C = \{c_1, \ldots, c_k\}$.

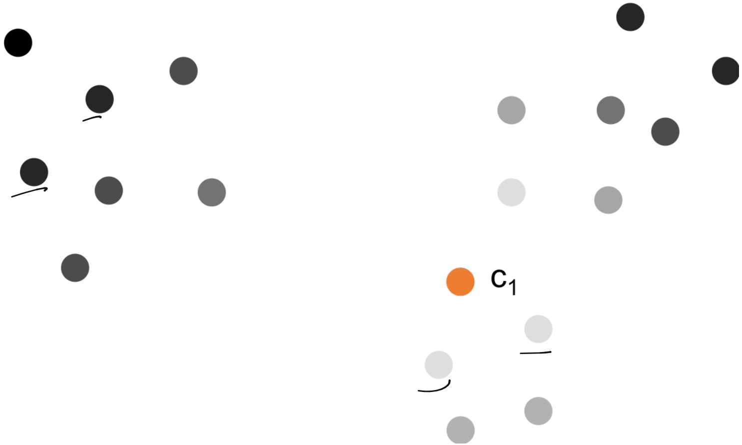
k-means++



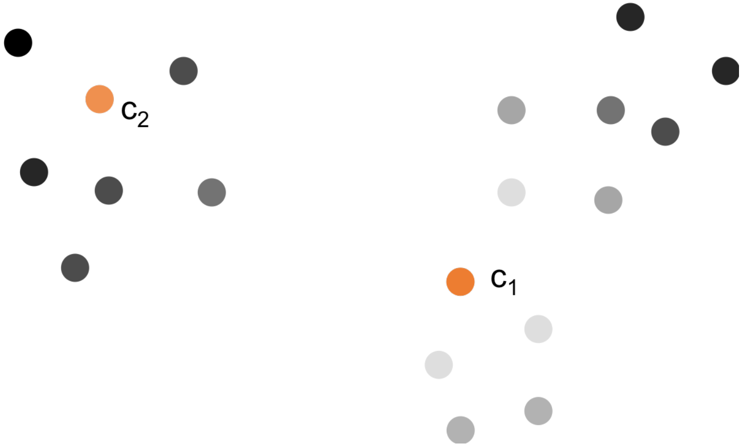
k-means++



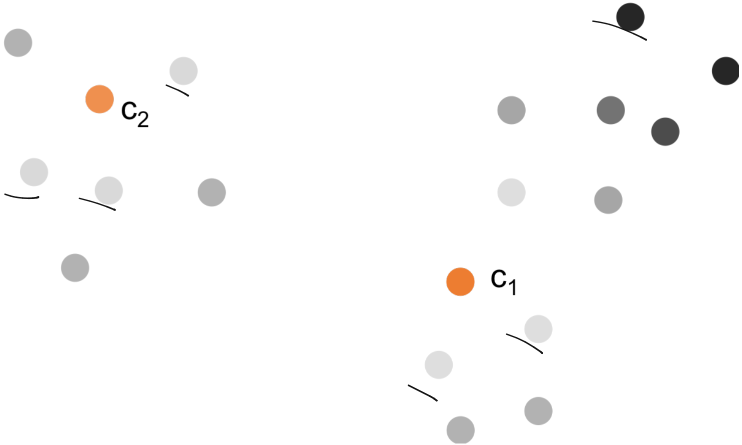
k-means++



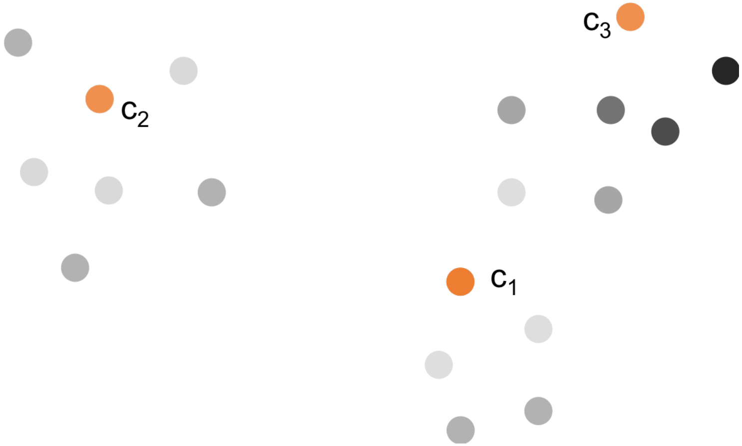
k-means++



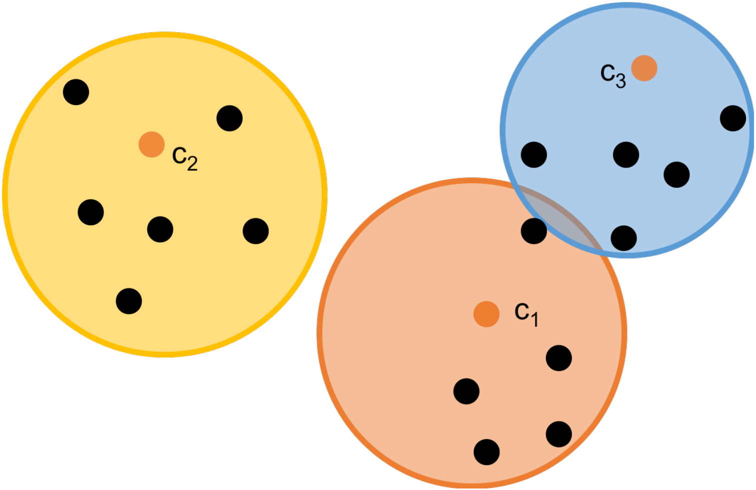
k-means++



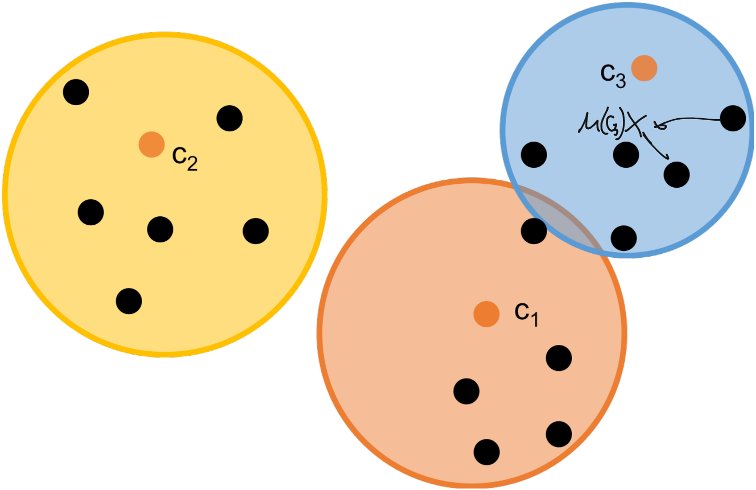
k-means++



k-means++



k-means++



Intuition: The adaptive sampling strategy tends to select well-spread cluster centers.

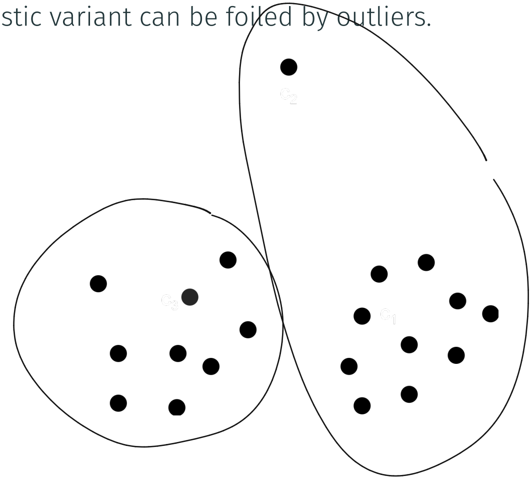
k-means++ Intuition

Why don't we just set c_j to the x_i with maximum $d_i = \min_{c \in C} \|x_i - c\|_2^2$? I.e., why do we use random sampling?

k-means++ Intuition

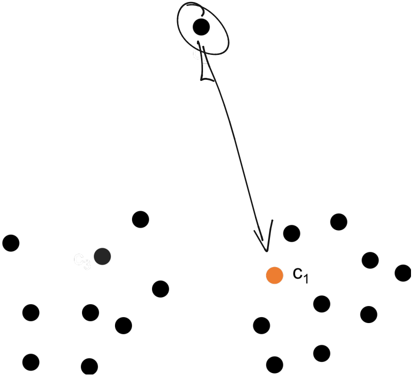
Why don't we just set c_j to the x_i with maximum $d_i = \min_{c \in C} \|x_i - c\|_2^2$? I.e., why do we use random sampling? This deterministic variant can be foiled by outliers.

$k = 2$

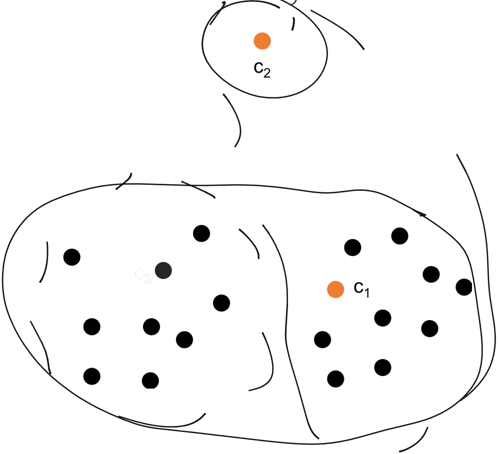


k-means++ Intuition

Why don't we just set c_j to the x_i with maximum $d_i = \min_{c \in C} \|x_i - c\|_2^2$? I.e., why do we use random sampling? This deterministic variant can be foiled by outliers.

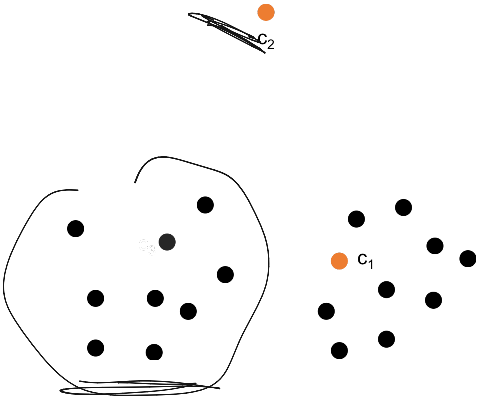


Why don't we just set c_j to the x_i with maximum $d_i = \min_{c \in C} \|x_i - c\|_2^2$? I.e., why do we use random sampling? This deterministic variant can be foiled by outliers.



k-means++ Intuition

Why don't we just set c_j to the x_i with maximum $d_i = \min_{c \in C} \|x_i - c\|_2^2$? I.e., why do we use random sampling? This deterministic variant can be foiled by outliers.



With random sampling cluster centers are both well-spread and representative of the dataset.

Proof Outline:

1. Let $C_1, \ldots, C_k$ the clusters corresponding to centers $c_1, \ldots, c_k$ and $\mu(C_1), \ldots, \mu(C_k)$ be their centroids. Let $A_1, \ldots, A_k$ be the optimal clusters. We will show:

$$\sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu(C_i)\|_2^2 \leq \sum_{i=1}^{k} \sum_{x \in C_i} \|x - c_i\|_2^2 \leq O(\log k) \cdot \sum_{i=1}^{k} \sum_{x \in A_i} \|x - \mu(A_i)\|_2^2.$$

c_i is a data point "representative" "seed" for cluster C_i

$\mu(C_i)$ centroid of that cluster

k-means++ Analysis

Proof Outline:

1. Let $C_1, \ldots, C_k$ the clusters corresponding to centers $c_1, \ldots, c_k$ and $\mu(C_1), \ldots, \mu(C_k)$ be their centroids. Let $A_1, \ldots, A_k$ be the optimal clusters. We will show:

$$\sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu(C_i)\|_2^2 \leq \sum_{i=1}^{k} \sum_{x \in C_i} \|x - c_i\|_2^2 \leq O(\log k) \cdot \sum_{i=1}^{k} \sum_{x \in A_i} \|x - \mu(A_i)\|_2^2.$$

2. Prove that, in expectation, the cost corresponding to any cluster A_i that has a center $c_1, \ldots, c_k$ selected from it (i.e., is covered) is at most a constant factor times the optimal cost.

k-means++ Analysis

Proof Outline:

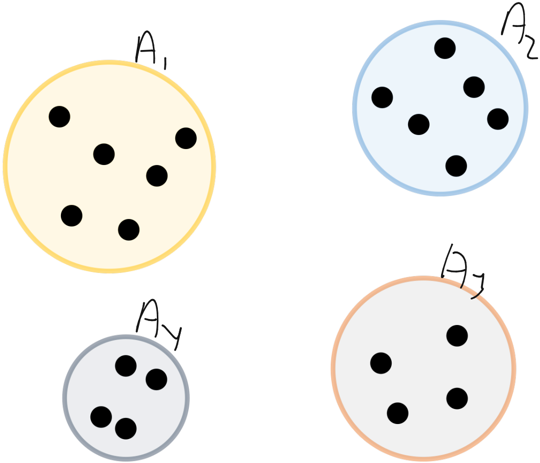
1. Let $C_1, \ldots, C_k$ the clusters corresponding to centers $c_1, \ldots, c_k$ and $\mu(C_1), \ldots, \mu(C_k)$ be their centroids. Let $A_1, \ldots, A_k$ be the optimal clusters. We will show:

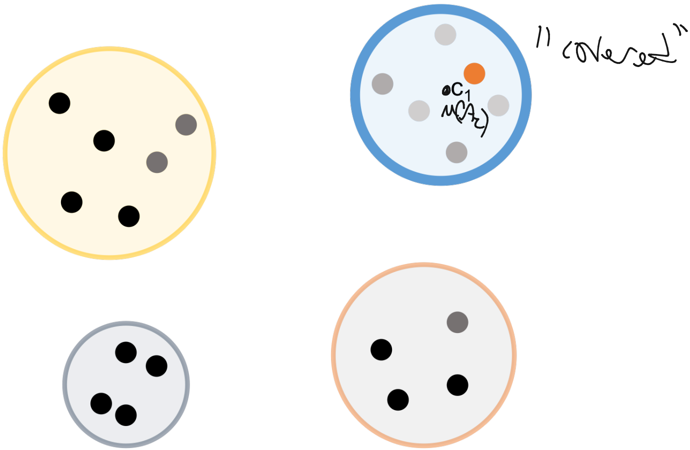
$$\sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu(C_i)\|_2^2 \leq \sum_{i=1}^{k} \sum_{x \in C_i} \|x - c_i\|_2^2 \leq O(\log k) \cdot \sum_{i=1}^{k} \sum_{x \in A_i} \|x - \mu(A_i)\|_2^2.$$

2. Prove that, in expectation, the cost corresponding to any cluster A_i that has a center $c_1, \ldots, c_k$ selected from it (i.e., is covered) is at most a constant factor times the optimal cost.
3. Argue that in each round of sampling, as long as the current cost is high, we are likely to select a new center from an uncovered cluster.

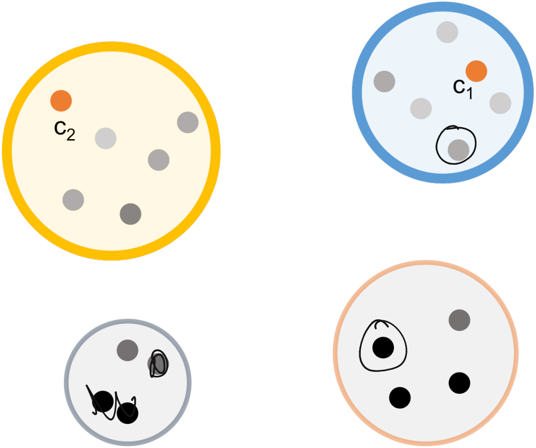
k-means++ Analysis

$k=4$

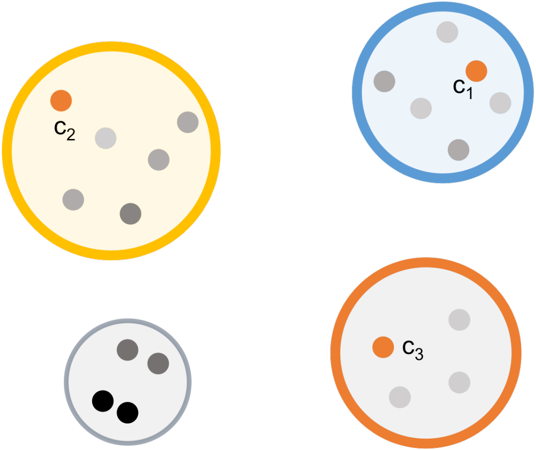



"covered"
C1

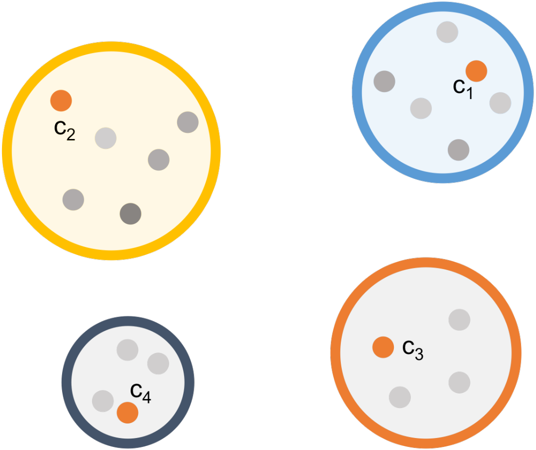
k-means++ Analysis



k-means++ Analysis



k-means++ Analysis



k-means++ Analysis

Proof Outline:

1. Let $C_1, \ldots, C_k$ the clusters corresponding to centers $c_1, \ldots, c_k$ and $\mu(C_1), \ldots, \mu(C_k)$ be their centroids. Let $A_1, \ldots, A_k$ be the optimal clusters. We will show:

$$\sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu(C_i)\|_2^2 \leq \sum_{i=1}^{k} \sum_{x \in C_i} \|x - c_i\|_2^2 \leq O(\log k) \cdot \sum_{i=1}^{k} \sum_{x \in A_i} \|x - \mu(A_i)\|_2^2.$$

2. Prove that, in expectation, the cost corresponding to any cluster A_i that has a center $c_1, \ldots, c_k$ selected from it (i.e., is covered) is at most a constant factor times the optimal cost.
3. Argue that in each round of sampling, as long as the current cost is high, we are likely to select a new center from an uncovered cluster.

k-means++ Analysis

Proof Outline:

1. Let $C_1, \ldots, C_k$ the clusters corresponding to centers $c_1, \ldots, c_k$ and $\mu(C_1), \ldots, \mu(C_k)$ be their centroids. Let $A_1, \ldots, A_k$ be the optimal clusters. We will show:

$$\sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu(C_i)\|_2^2 \leq \sum_{i=1}^{k} \sum_{x \in C_i} \|x - c_i\|_2^2 \leq O(\log k) \cdot \sum_{i=1}^{k} \sum_{x \in A_i} \|x - \mu(A_i)\|_2^2.$$

2. Prove that, in expectation, the cost corresponding to any cluster A_i that has a center $c_1, \ldots, c_k$ selected from it (i.e., is covered) is at most a constant factor times the optimal cost.
3. Argue that in each round of sampling, as long as the current cost is high, we are likely to select a new center from an uncovered cluster.
4. Conclude that we cover any high cost clusters with good probability, and via a careful inductive argument that the expected cost is $O(\log k)$ times the optimum.

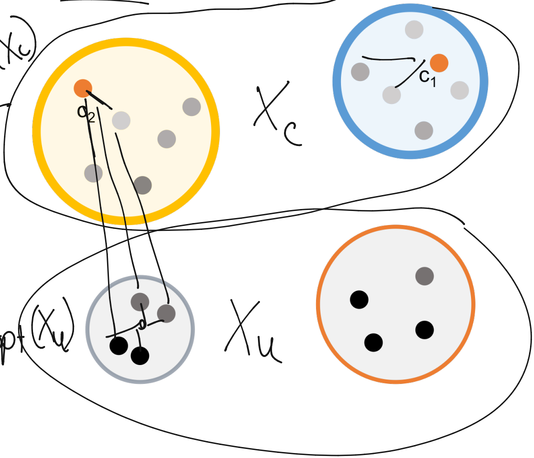
Proof Outline:

1. Let $C_1, \ldots, C_k$ the clusters corresponding to centers $c_1, \ldots, c_k$ and $\mu(C_1), \ldots, \mu(C_k)$ be their centroids. Let $A_1, \ldots, A_k$ be the optimal clusters. We will show:
$$\sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu(C_i)\|_2^2 \leq \sum_{i=1}^{k} \sum_{x \in C_i} \|x - c_i\|_2^2 \leq O(\log k) \cdot \sum_{i=1}^{k} \sum_{x \in A_i} \|x - \mu(A_i)\|_2^2.$$
2. Prove that, in expectation, the cost corresponding to any cluster A_i that has a center $c_1, \ldots, c_k$ selected from it (i.e., is covered) is at most a constant factor times the optimal cost.
3. Argue that in each round of sampling, as long as the current cost is high, we are likely to select a new center from an uncovered cluster.
4. Conclude that we cover any high cost clusters with good probability, and via a careful inductive argument that the expected cost is $O(\log k)$ times the optimum.

k-means++ Proof Sketch

Let $\mathcal{X}_u, \mathcal{X}_c$ be the set of uncovered and covered points respectively. $\mathcal{X} = X_u \cup X_c$
Let $\phi(\mathcal{X}_u)$ and $\phi(\mathcal{X}_c)$ be the current cost associated with these points, and $\phi_{OPT}(\mathcal{X}_u)$ and $\phi_{OPT}(\mathcal{X}_c)$ denote the optimal cost.

$\phi(X_c) \simeq \phi_{opt}(X_c)$



$\phi(X_u) >> \phi_{opt}(X_u)$

k-means++ Proof Sketch

Let $\mathcal{X}_u, \mathcal{X}_c$ be the set of uncovered and covered points respectively. Let $\phi(\mathcal{X}_u)$ and $\phi(\mathcal{X}_c)$ be the current cost associated with these points, and $\phi_{OPT}(\mathcal{X}_u)$ and $\phi_{OPT}(\mathcal{X}_c)$ denote the optimal cost.

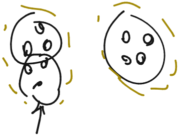
- Will argue in a few slides that $\mathbb{E}[\phi(\mathcal{X}_c)] \overset{8}{\lesssim} \phi_{OPT}(\mathcal{X}_c)$

k-means++ Proof Sketch

Let $\mathcal{X}_u, \mathcal{X}_c$ be the set of uncovered and covered points respectively. Let $\phi(\mathcal{X}_u)$ and $\phi(\mathcal{X}_c)$ be the current cost associated with these points, and $\phi_{OPT}(\mathcal{X}_u)$ and $\phi_{OPT}(\mathcal{X}_c)$ denote the optimal cost.

- Will argue in a few slides that $\mathbb{E}[\phi(\mathcal{X}_c)] \lesssim \phi_{OPT}(\mathcal{X}_c)$
- If $\phi(\mathcal{X}) \geq \alpha \cdot \phi_{OPT}(\mathcal{X})$, then $\phi(\mathcal{X}_c) \leq \phi_{OPT}(\mathcal{X}_c) \leq \phi_{OPT}(\mathcal{X}) \leq 1/\alpha \cdot \phi(\mathcal{X})$.

$k=3$



k-means++ Proof Sketch

Let $\mathcal{X}_u, \mathcal{X}_c$ be the set of uncovered and covered points respectively. Let $\phi(\mathcal{X}_u)$ and $\phi(\mathcal{X}_c)$ be the current cost associated with these points, and $\phi_{OPT}(\mathcal{X}_u)$ and $\phi_{OPT}(\mathcal{X}_c)$ denote the optimal cost.

- Will argue in a few slides that $\mathbb{E}[\phi(\mathcal{X}_c)] \lesssim \phi_{OPT}(\mathcal{X}_c)$
- If $\phi(\mathcal{X}) \geq \alpha \cdot \phi_{OPT}(\mathcal{X})$, then $\phi(\mathcal{X}_c) \leq \phi_{OPT}(\mathcal{X}_c) \leq \phi_{OPT}(\mathcal{X}) \leq 1/\alpha \cdot \phi(\mathcal{X})$. So $\phi(\mathcal{X}_u) = \phi(\mathcal{X}) - \phi(\mathcal{X}_c) \geq (1 - 1/\alpha) \cdot \phi(\mathcal{X})$. So, we cover a new cluster with probability:
$$\frac{\phi(\mathcal{X}_u)}{\phi(\mathcal{X})} \geq 1 - \frac{1}{\alpha} \approx 1$$
when α is large.
- I.e., unless our current cost is close to the optimal cost, we cover a new cluster with high probability in each step.

k-means++ Analysis

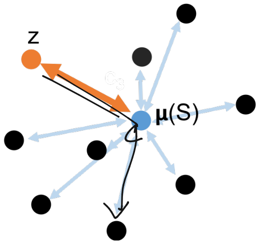
k-means++ Analysis

It remains to show that in expectation, the cost corresponding to a covered cluster A_i is at most a constant factor times the optimal cost.

It remains to show that in expectation, the cost corresponding to a covered cluster A_i is at most a constant factor times the optimal cost.

A Useful Lemma: Let S be a set of points with centroid $\mu(S)$, and let z be any other point.

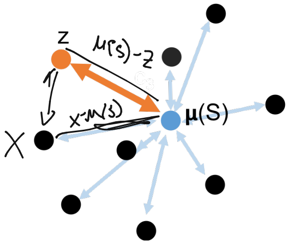
$$\sum_{x \in S} \|x - z\|_2^2 = \sum_{x \in S} \|x - \mu(S)\|_2^2 + |S| \cdot \|\mu(S) - z\|_2^2.$$



It remains to show that in expectation, the cost corresponding to a covered cluster A_i is at most a constant factor times the optimal cost.

A Useful Lemma: Let S be a set of points with centroid $\mu(S)$, and let z be any other point.

$$\sum_{x \in S} \|x - z\|_2^2 = \sum_{x \in S} \|x - \mu(S)\|_2^2 + |S| \cdot \|\mu(S) - z\|_2^2.$$



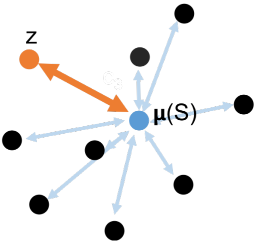
Proof: $\sum_{x \in S} \|x - z\|_2^2 = \sum_{x \in S} \|(x - \mu(S)) + (\mu(S) - z)\|_2^2$

k-means++ Analysis

It remains to show that in expectation, the cost corresponding to a covered cluster A_i is at most a constant factor times the optimal cost.

A Useful Lemma: Let S be a set of points with centroid $\mu(S)$, and let z be any other point.

$$\sum_{x\in S}\|x-z\|_2^2 = \overbrace{\sum_{x\in S}\|x-\mu(S)\|_2^2}^{1} + \overbrace{|S|\cdot\|\mu(S)-z\|_2^2}^{2}.$$



$$\sum_{x\in S} x - \frac{1}{|S|}\sum_{x\in S} x$$

$$\sum_{x\in S} x - \sum_{x\in S} x = 0$$

$$\sum_{x\in S}\langle x-\mu(S), \mu(S)-z\rangle$$

$$\langle \underbrace{\sum_{x\in S} x-\mu(S)}_{0}, \mu(S)-z\rangle$$

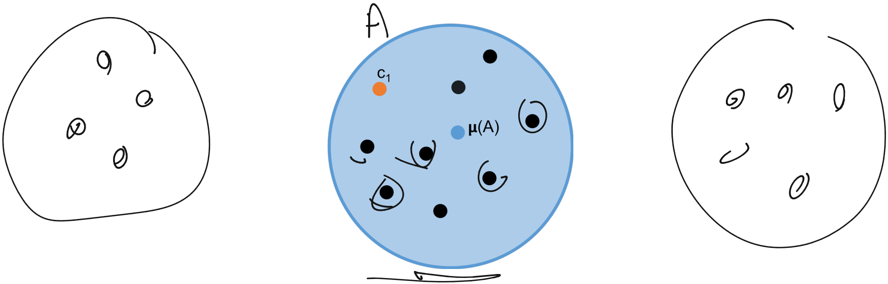
Proof: $\sum_{x\in S}\|x-z\|_2^2 = \sum_{x\in S}\|(x-\mu(S))+(\mu(S)-z)\|_2^2 =$
$\underbrace{\sum_{x\in S}\|x-\mu(S)\|_2^2}_{1} + \underbrace{\sum_{x\in S}\|\mu(S)-z\|_2^2}_{2} + \sum_{x\in S}2\langle x-\mu(S),\mu(S)-z\rangle$

First Cluster Bound

Lemma

Let A be some cluster in the optimal cluster set $A_1, \ldots, A_k$. Let c_1 be a cluster center chosen uniformly at random from A. Let $\phi(A) = \sum_{x \in A} \|x - c_1\|_2^2$ and $\phi_{OPT}(A) = \sum_{x \in A} \|x - \mu(A)\|_2^2$.

$$\mathbb{E}[\phi(A)] = 2\phi_{OPT}(A).$$



First Cluster Bound

Lemma

Let A be some cluster in the optimal cluster set $A_1, \ldots, A_k$. Let c_1 be a cluster center chosen uniformly at random from A. Let $\phi(A) = \sum_{x \in A} \|x - c_1\|_2^2$ and $\phi_{OPT}(A) = \sum_{x \in A} \|x - \mu(A)\|_2^2$.

$$\mathbb{E}[\phi(A)] = 2\phi_{OPT}(A).$$

$$\mathbb{E}[\phi(A)] = \sum_{a_1 \in A} \frac{1}{|A|} \cdot \sum_{a_2 \in A} \|a_1 - a_2\|_2^2$$

First Cluster Bound

Lemma

Let A be some cluster in the optimal cluster set $A_1, \ldots, A_k$. Let c_1 be a cluster center chosen uniformly at random from A. Let $\phi(A) = \sum_{x \in A} \|x - c_1\|_2^2$ and $\phi_{OPT}(A) = \sum_{x \in A} \|x - \mu(A)\|_2^2$.

$$\mathbb{E}[\phi(A)] = 2\phi_{OPT}(A).$$

$$\mathbb{E}[\phi(A)] = \sum_{a_1 \in A} \frac{1}{|A|} \cdot \sum_{a_2 \in A} \|a_1 - a_2\|_2^2$$

$$= \frac{1}{|A|} \sum_{a_1 \in A} \sum_{a_2 \in A} \left[\|a_2 - \mu(A)\|_2^2 + |A| \cdot \|a_1 - \mu(A)\|_2^2 \right]$$

$$\cancel{\frac{1}{|A|}} \cdot |A| \cdot \underbrace{\sum_{a_2 \in A} \|a_2 - \mu(A)\|_2^2}_{\phi_{OPT}(A)} + \cancel{\frac{1}{|A|}} |A| \sum_{a_1 \in A} \|a_1 - \mu(A)\|_2^2$$

First Cluster Bound

Lemma

Let A be some cluster in the optimal cluster set $A_1, \ldots, A_k$. Let c_1 be a cluster center chosen uniformly at random from A. Let $\phi(A) = \sum_{x \in A} \|x - c_1\|_2^2$ and $\phi_{OPT}(A) = \sum_{x \in A} \|x - \mu(A)\|_2^2$.

$$\mathbb{E}[\phi(A)] = 2\phi_{OPT}(A).$$

$$\begin{aligned}
\mathbb{E}[\phi(A)] &= \sum_{a_1 \in A} \frac{1}{|A|} \cdot \sum_{a_2 \in A} \|a_1 - a_2\|_2^2 \\
&= \frac{1}{|A|} \sum_{a_1 \in A} \sum_{a_2 \in A} \left[\|a_2 - \mu(A)\|_2^2 + |A| \cdot \|a_1 - \mu(A)\|_2^2 \right] \\
&= \sum_{a_1 \in A} \|a_1 - \mu(A)\|_2^2 + \sum_{a_2 \in A} \|a_2 - \mu(A)\|_2^2
\end{aligned}$$

First Cluster Bound

Lemma

Let A be some cluster in the optimal cluster set $A_1, \ldots, A_k$. Let c_1 be a cluster center chosen uniformly at random from A. Let $\phi(A) = \sum_{x \in A} \|x - c_1\|_2^2$ and $\phi_{OPT}(A) = \sum_{x \in A} \|x - \mu(A)\|_2^2$.

$$\mathbb{E}[\phi(A)] = 2\phi_{OPT}(A).$$

$$\begin{aligned}
\mathbb{E}[\phi(A)] &= \sum_{a_1 \in A} \frac{1}{|A|} \cdot \sum_{a_2 \in A} \|a_1 - a_2\|_2^2 \\
&= \frac{1}{|A|} \sum_{a_1 \in A} \sum_{a_2 \in A} \left[\|a_2 - \mu(A)\|_2^2 + |A| \cdot \|a_1 - \mu(A)\|_2^2\right] \\
&= \sum_{a_1 \in A} \|a_1 - \mu(A)\|_2^2 + \sum_{a_2 \in A} \|a_2 - \mu(A)\|_2^2 \\
&= 2\phi_{OPT}(A).
\end{aligned}$$

Future Cluster Bounds

Lemma

Let A be some cluster in the optimal cluster set $A_1, \ldots, A_k$. Let $c_1, \ldots, c_{j-1}$ be our current set of cluster centers. If we add a random center c_j from A, chosen with probability proportional to $d(a) = \min_{i \in \{1,\ldots,j-1\}} \|a - c_i\|_2^2$ then

$$\frac{d(a)}{\sum_{a \in A} d(a)}$$

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

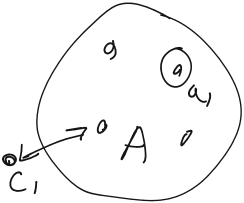
Future Cluster Bounds

Lemma

Let A be some cluster in the optimal cluster set $A_1, \ldots, A_k$. Let $c_1, \ldots, c_{j-1}$ be our current set of cluster centers. If we add a random center c_j from A, chosen with probability proportional to $d(a) = \min_{i \in \{1,\ldots,j-1\}} \|a - c_i\|_2^2$ then

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

$$\underline{\mathbb{E}[\phi(A)]} = \sum_{a_1 \in A} \frac{d(a_1)}{\sum_{a \in A} d(a)} \cdot \sum_{a_2 \in A} \min(\underline{d(a_2)}, \underline{\|a_2 - a_1\|_2^2})$$

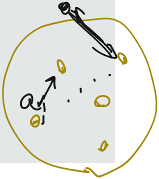


Future Cluster Bounds

Lemma

Let A be some cluster in the optimal cluster set $A_1, \ldots, A_k$. Let $c_1, \ldots, c_{j-1}$ be our current set of cluster centers. If we add a random center c_j from A, chosen with probability proportional to $d(a) = \min_{i \in \{1,\ldots,j-1\}} \|a - c_i\|_2^2$ then

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



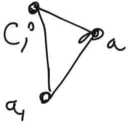
$$\mathbb{E}[\phi(A)] = \sum_{a_1 \in A} \frac{d(a_1)}{\sum_{a \in A} d(a)} \cdot \sum_{a_2 \in A} \min(d(a_2), \|a_2 - a_1\|_2^2)$$

By triangle inequality, for any center c_i, a_1 , a any points

$\|a_1 - c_i\|_2^2 \leq (\|a - c_i\|_2 + \|a - a_1\|_2)^2 \leq 2\|a - c_i\|_2^2 + 2\|a - a_1\|_2^2$. So

Δ inequality

$$d(a_1) \leq 2d(a) + 2\|a - a_1\|_2^2.$$



$d(a) = \|c_i - a\|_2^2$ $\quad d(A)$

$d(a_1) \leq \|c_i - a_1\|_2^2 \leq 2\|a - c_i\|^2 + 2\|a - a_1\|_2^2$

Future Cluster Bounds

Lemma

Let A be some cluster in the optimal cluster set $A_1, \ldots, A_k$. Let $c_1, \ldots, c_{j-1}$ be our current set of cluster centers. If we add a random center c_j from A, chosen with probability proportional to $d(a) = \min_{i \in \{1,\ldots,j-1\}} \|a - c_i\|_2^2$ then

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

$$\mathbb{E}[\phi(A)] = \sum_{a_1 \in A} \frac{d(a_1)}{\sum_{a \in A} d(a)} \cdot \sum_{a_2 \in A} \min(d(a_2), \|a_2 - a_1\|_2^2)$$

By triangle inequality, for any center c_i, a_1 a_1

$\|a_1 - c_i\|_2^2 \leq (\|a - c_i\|_2 + \|a - a_1\|_2)^2 \leq 2\|a - c_i\|_2^2 + 2\|a - a_1\|_2^2$. So

$$d(a_1) \leq 2d(a) + 2\|a - a_1\|_2^2.$$

Averaging over all $a \in A$, $d(a_1) \leq \frac{2}{|A|} \sum_{a \in A} d(a) + \frac{2}{|A|} \sum_{a \in A} \|a - a_1\|_2^2$.

Combine: $\mathbb{E}[\phi(A)] = \sum_{a_1 \in A} \frac{d(a_1)}{\sum_{a \in A} d(a)} \cdot \sum_{a_2 \in A} \min(d(a_2), \|a_2 - a_1\|_2^2)$
and $d(a_1) \leq \frac{2}{|A|} \sum_{a \in A} d(a) + \frac{2}{|A|} \sum_{a \in A} \|a - a_1\|_2^2$ to get:

$$\mathbb{E}[\phi(A)] \leq \frac{2}{|A|} \left(\sum_{a_1 \in A} \frac{\sum_{a \in A} d(a)}{\sum_{a \in A} d(a)} \sum_{a_2 \in A} \|a_2 - a_1\|_2^2 + \sum_{a_1 \in A} \frac{\sum_{a \in A} \|a - a_1\|_2^2}{\sum_{a \in A} d(a)} \sum_{a_2 \in A} d(a_2) \right)$$

Future Cluster Bounds

Combine: $\mathbb{E}[\phi(A)] = \sum_{a_1 \in A} \frac{d(a_1)}{\sum_{a \in A} d(a)} \cdot \sum_{a_2 \in A} \min(d(a_2), \|a_2 - a_1\|_2^2)$
and $d(a_1) \leq \frac{2}{|A|} \sum_{a \in A} d(A) + \frac{2}{|A|} \sum_{a \in A} \|a - a_1\|_2^2$ to get:

$$
\begin{aligned}
\mathbb{E}[\phi(A)] &\leq \frac{2}{|A|} \left(\sum_{a_1 \in A} \frac{\sum_{a \in A} d(A)}{\sum_{a \in A} d(A)} \sum_{a_2 \in A} \|a_2 - a_1\|_2^2 + \sum_{a_1 \in A} \frac{\sum_{a \in A} \|a - a_1\|_2^2}{\cancel{\sum_{a \in A} d(A)}} \cancel{\sum_{a_2 \in A} d(a_2)} \right) \\
&= \frac{4}{|A|} \sum_{a_1 \in A} \sum_{a_2 \in A} \|a_2 - a_1\|_2^2
\end{aligned}
$$

Future Cluster Bounds

Combine: $\mathbb{E}[\phi(A)] = \sum_{a_1 \in A} \frac{d(a_1)}{\sum_{a \in A} d(a)} \cdot \sum_{a_2 \in A} \min(d(a_2), \|a_2 - a_1\|_2^2)$

and $d(a_1) \leq \frac{2}{|A|} \sum_{a \in A} d(A) + \frac{2}{|A|} \sum_{a \in A} \|a - a_1\|_2^2$ to get:

$$\mathbb{E}[\phi(A)] \leq \frac{2}{|A|} \left(\sum_{a_1 \in A} \frac{\sum_{a \in A} d(A)}{\sum_{a \in A} d(A)} \sum_{a_2 \in A} \|a_2 - a_1\|_2^2 + \sum_{a_1 \in A} \frac{\sum_{a \in A} \|a - a_1\|_2^2}{\sum_{a \in A} d(A)} \sum_{a_2 \in A} d(a_2) \right)$$

$$= \frac{4}{|A|} \underbrace{\sum_{a_1 \in A} \sum_{a_2 \in A} \|a_2 - a_1\|_2^2}_{2\phi_{opt}(A)} \leq 8\phi_{OPT}(A).$$

Future Cluster Bounds

Combine: $\mathbb{E}[\phi(A)] = \sum_{a_1 \in A} \frac{d(a_1)}{\sum_{a \in A} d(a)} \cdot \sum_{a_2 \in A} \min(d(a_2), \|a_2 - a_1\|_2^2)$
and $d(a_1) \le \frac{2}{|A|} \sum_{a \in A} d(A) + \frac{2}{|A|} \sum_{a \in A} \|a - a_1\|_2^2$ to get:

$$\mathbb{E}[\phi(A)] \le \frac{2}{|A|} \left(\sum_{a_1 \in A} \frac{\sum_{a \in A} d(A)}{\sum_{a \in A} d(A)} \sum_{a_2 \in A} \|a_2 - a_1\|_2^2 + \sum_{a_1 \in A} \frac{\sum_{a \in A} \|a - a_1\|_2^2}{\sum_{a \in A} d(A)} \sum_{a_2 \in A} d(a_2) \right)$$
$$= \frac{4}{|A|} \sum_{a_1 \in A} \sum_{a_2 \in A} \|a_2 - a_1\|_2^2 \le 8\phi_{OPT}(A).$$

Upshot: At each step that we cover a cluster A from the optimal clustering, the expected cost is, in expectation, within a constant factor of the optimal cost for that cluster.

Randomized Low-Rank approximation

Consider a matrix $A \in \mathbb{R}^{n \times d}$. We would like to compute an optimal low-rank approximation of A. I.e., for $k \ll \min(n, d)$ we would like to find $Z \in \mathbb{R}^{n \times k}$ with orthonormal columns satisfying:

$Z^T Z = I$

$$\|\underline{A - ZZ^T A}\|_F = \min_{\underbrace{Z:Z^T Z=I}_{Z \text{ is orthonormal}}} \|A - ZZ^T A\|_F.$$

ZZ^T is the projection onto Z's span

Low-rank Approximation

Consider a matrix $A \in \mathbb{R}^{n \times d}$. We would like to compute an optimal low-rank approximation of A. I.e., for $k \ll \min(n, d)$ we would like to find $Z \in \mathbb{R}^{n \times k}$ with orthonormal columns satisfying:

$$\|A - ZZ^T A\|_F = \min_{Z : Z^T Z = I} \|A - ZZ^T A\|_F.$$

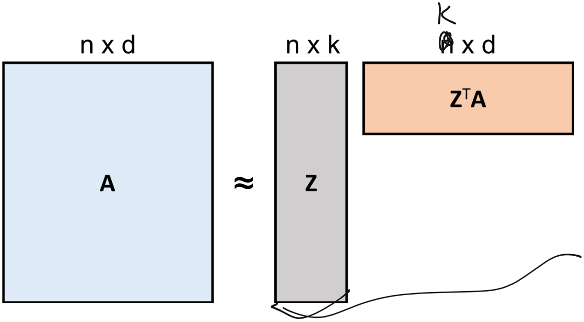
Why is $\text{rank}(ZZ^T A) \leq k$?

Low-rank Approximation

Consider a matrix $A \in \mathbb{R}^{n \times d}$. We would like to compute an optimal low-rank approximation of A. I.e., for $k \ll \min(n, d)$ we would like to find $Z \in \mathbb{R}^{n \times k}$ with orthonormal columns satisfying:

$$\|A - ZZ^T A\|_F = \min_{Z : Z^T Z = I} \|A - ZZ^T A\|_F.$$

Why is $\text{rank}(ZZ^T A) \leq k$?



Low-rank Approximation

Consider a matrix $A \in \mathbb{R}^{n \times d}$. We would like to compute an optimal low-rank approximation of A. I.e., for $k \ll \min(n, d)$ we would like to find $Z \in \mathbb{R}^{n \times k}$ with orthonormal columns satisfying:

$$\|A - ZZ^T A\|_F = \min_{Z:Z^T Z = I} \|A - ZZ^T A\|_F.$$

$$\min_{B:\,\mathrm{rank}(B)=k} \|A - B\|_F$$

Why is $\mathrm{rank}(ZZ^T A) \leq k$?

Why does it suffice to consider low-rank approximations of this form?

Low-rank Approximation

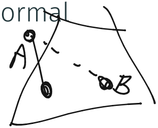
Consider a matrix $A \in \mathbb{R}^{n\times d}$. We would like to compute an optimal low-rank approximation of A. I.e., for $k \ll \min(n,d)$ we would like to find $Z \in \mathbb{R}^{n\times k}$ with orthonormal columns satisfying:

$$\|A - ZZ^TA\|_F = \min_{Z:Z^TZ=I} \|A - ZZ^TA\|_F.$$

Why is $\text{rank}(ZZ^TA) \le k$?

$n\times k$ $k\times d$
$M \cdot N$

Why does it suffice to consider low-rank approximations of this form? For any B with $\text{rank}(B) = k$, let $Z \in \mathbb{R}^{n\times k}$ be an orthonormal basis for B's column span. Then $\|A - ZZ^TA\|_F \le \|A - B\|_F$. So



$$\min_{Z:Z^TZ=I} \|A - ZZ^TA\|_F = \min_{B:\text{rank } B=k} \|A - B\|_F.$$

Low-rank Approximation

Consider a matrix $A \in \mathbb{R}^{n \times d}$. We would like to compute an optimal low-rank approximation of A. I.e., for $k \ll \min(n, d)$ we would like to find $Z \in \mathbb{R}^{n \times k}$ with orthonormal columns satisfying:

$$\|A - ZZ^T A\|_F = \min_{Z: Z^T Z = I} \|A - ZZ^T A\|_F.$$

Why is $\text{rank}(ZZ^T A) \leq k$?

Why does it suffice to consider low-rank approximations of this form? For any B with $\text{rank}(B) = k$, let $Z \in \mathbb{R}^{n \times k}$ be an orthonormal basis for B's column span. Then $\|A - ZZ^T A\|_F \leq \|A - B\|_F$. So

$$\min_{Z: Z^T Z = I} \|A - ZZ^T A\|_F = \min_{B: \text{rank } B = k} \|A - B\|_F.$$

How would one compute the optimal basis Z?

Consider a matrix $A \in \mathbb{R}^{n \times d}$. We would like to compute an optimal low-rank approximation of A. I.e., for $k \ll \min(n, d)$ we would like to find $Z \in \mathbb{R}^{n \times k}$ with orthonormal columns satisfying:

$\|A - ZZ^T A\|_F^2$
$= tr(AA^T) - \underline{tr(Z^T A A^T Z)}$

$$\|A - ZZ^T A\|_F = \min_{Z: Z^T Z = I} \|A - ZZ^T A\|_F.$$

Why is $\text{rank}(ZZ^T A) \leq k$?

Courant fischer theorem

Why does it suffice to consider low-rank approximations of this form? For any B with $\text{rank}(B) = k$, let $Z \in \mathbb{R}^{n \times k}$ be an orthonormal basis for B's column span. Then $\|A - ZZ^T A\|_F \leq \|A - B\|_F$. So

$$\min_{Z: Z^T Z = I} \|A - ZZ^T A\|_F = \min_{B: \text{rank } B = k} \|A - B\|_F.$$

$AA^T =$
$U \Sigma^2 U^T$

How would one compute the optimal basis Z? Compute the top k left singular vectors of A, which requires $O(nd^2)$ time, or $\boxed{O(ndk)}$ time for a high accuracy approximation with an iterative method.

$A = \underset{n \times r}{U} \underset{r \times r}{\Sigma} \underset{r \times d}{V^T}$

$n \times r$

Sampling Based Algorithm

We will analysis a simple non-uniform sampling based algorithm for low-rank approximation, that gives a near optimal solution in $O(nd + nk^2)$ time. $O(nd^2)$ $O(ndk)$

Sampling Based Algorithm

We will analysis a simple non-uniform sampling based algorithm for low-rank approximation, that gives a near optimal solution in $O(\underline{nd} + nk^2)$ time.

Linear Time Low-Rank Approximation:

pick t colums of A

- Fix sampling probabilities $p_1, \ldots, p_n$ with $\underline{p_i} = \underline{\frac{\|A_{:,i}\|_2^2}{\|A\|_F^2}}$.
- Select $\mathbf{i_1}, \ldots, \mathbf{i_t} \in [n]$ independently, according to the distribution $\Pr[\mathbf{i_j} = k] = p_k$ for sample size $t \geq k$.
- $\cancel{\text{Let } \mathbf{C} = \frac{1}{t} \cdot \sum_{j=1}^{t} \frac{1}{\sqrt{p_{\mathbf{i_j}}}} \cdot A_{:,\mathbf{i_j}}}$. $C \in \mathbb{R}^{n \times t}$ $C_{:j} = \frac{1}{\sqrt{t p_{i_j}}} \cdot A_{:i_j}$
- Let $\overline{\mathbf{Z}} \in \mathbb{R}^{n \times k}$ consist of the top k left singular vectors of $\mathbf{C}$.

We will analysis a simple non-uniform sampling based algorithm for low-rank approximation, that gives a near optimal solution in $O(nd + nk^2)$ time.

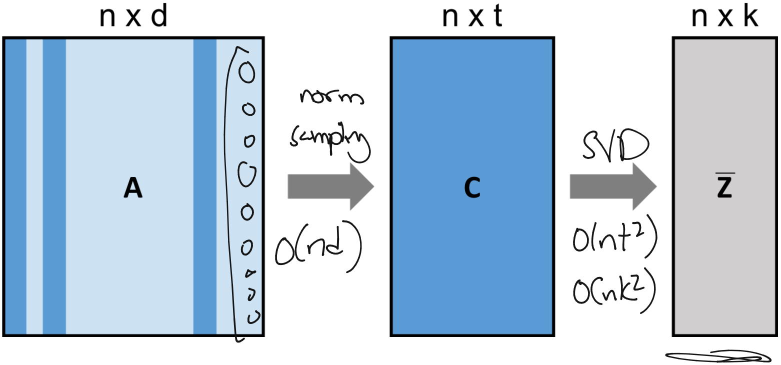
Linear Time Low-Rank Approximation:

- Fix sampling probabilities $p_1, \ldots, p_n$ with $p_i = \frac{\|A_{:,i}\|_2^2}{\|A\|_F^2}$.
- Select $\mathbf{i_1}, \ldots, \mathbf{i_t} \in [n]$ independently, according to the distribution $\Pr[\mathbf{i_j} = k] = p_k$ for sample size $t \geq k$.
- Let $\mathbf{C} = \frac{1}{t} \cdot \sum_{j=1}^{t} \frac{1}{\sqrt{p_{\mathbf{i_j}}}} \cdot A_{:,\mathbf{i_j}}$.
- Let $\bar{\mathbf{Z}} \in \mathbb{R}^{n \times k}$ consist of the top k left singular vectors of $\mathbf{C}$.

Looks like approximate matrix multiplication! In fact, will use that $\mathbf{C}\mathbf{C}^T$ is a good approximation to the matrix product AA^T.

Sampling Based Algorithm

$t \approx k$



$$\|A - \underline{Z}Z^T A\|_F$$

Sampling Based Algorithm Approximation Bound

Theorem

The linear time low-rank approximation algorithm run with $t = \frac{k}{\epsilon^2 \cdot \sqrt{\delta}}$ samples outputs $\bar{\mathbf{Z}} \in \mathbb{R}^{n\times k}$ satisfying with probability at least $1-\delta$:

SVD

$$\|A - \bar{\mathbf{Z}}\bar{\mathbf{Z}}^T A\|_F^2 \leq \min_{Z:Z^TZ=I} \|A - ZZ^TA\|_F^2 + 2\epsilon\|A\|_F^2.$$

Sampling Based Algorithm Approximation Bound

Theorem

The linear time low-rank approximation algorithm run with $t = \frac{k}{\epsilon^2 \cdot \sqrt{\delta}}$ samples outputs $\bar{\mathbf{Z}} \in \mathbb{R}^{n \times k}$ satisfying with probability at least $1 - \delta$:

$$\|A - \bar{\mathbf{Z}}\bar{\mathbf{Z}}^T A\|_F^2 \le \min_{Z:Z^TZ=I} \|A - ZZ^TA\|_F^2 + 2\epsilon\|A\|_F^2.$$

Key Idea: By the approximate matrix multiplication result from last class, applied to the matrix product AA^T, with probability $\geq 1 - \delta$,

$$\|AA^T - \mathbf{C}\mathbf{C}^T\|_F \le \frac{\epsilon}{\sqrt{k}} \cdot \|A\|_F \cdot \|A^T\|_F = \frac{\epsilon}{\sqrt{k}}\|A\|_F^2.$$

Sampling Based Algorithm Approximation Bound

Theorem

Could I sample << k columns.

The linear time low-rank approximation algorithm run with $t = \cancel{\frac{k}{\epsilon^2 \cdot \sqrt{\delta}}}$ samples outputs $\bar{\mathbf{Z}} \in \mathbb{R}^{n \times k}$ satisfying with probability at least $1 - \delta$:

$\frac{1}{\epsilon^2}$ $\epsilon \|A\|_F \|B\|_F$

$$\|A - \bar{\mathbf{Z}}\bar{\mathbf{Z}}^T A\|_F^2 \leq \min_{Z: Z^T Z = I} \|A - ZZ^T A\|_F^2 + 2\epsilon \|A\|_F^2.$$

Key Idea: By the approximate matrix multiplication result from last class, applied to the matrix product AA^T, with probability $\geq 1 - \delta$,

$$\|AA^T - \mathbf{C}\mathbf{C}^T\|_F \leq \frac{\epsilon}{\sqrt{k}} \cdot \|A\|_F \cdot \|A^T\|_F = \frac{\epsilon}{\sqrt{k}} \|A\|_F^2.$$

$\text{rank}(CC^T) < k$

Since $\mathbf{C}\mathbf{C}^T$ is close to AA^T, the top eigenvectors of these matrices (i.e. the top left singular vectors of A and $\mathbf{C}$ will not be too different.) So $\bar{\mathbf{Z}}$ can be used in place of the top left singular vectors of A to give a near optimal approximation.

Formal Analysis

Let $Z_* \in \mathbb{R}^{n \times k}$ contain the top left singular vectors of A – i.e. $Z_* = \arg\min \|A - ZZ^TA\|_F^2$. Similarly, $\bar{\mathbf{Z}} = \arg\min \|\mathbf{C} - ZZ^T\mathbf{C}\|_F^2$.

Formal Analysis

Let $Z_* \in \mathbb{R}^{n\times k}$ contain the top left singular vectors of A – i.e. $Z_* = \arg\min \|A - ZZ^TA\|_F^2$. Similarly, $\bar{\mathbf{Z}} = \arg\min \|\mathbf{C} - ZZ^T\mathbf{C}\|_F^2$.

Claim 1: For any orthonormal $Z \in \mathbb{R}^{n\times k}$, and any matrix B,

$$\|B - ZZ^TB\|_F^2 = \mathrm{tr}(BB^T) - \mathrm{tr}(Z^TBB^TZ).$$

$$\mathrm{tr}\left[(B - ZZ^TB)(B - ZZ^TB)^T\right]$$

$$\mathrm{tr}(BB^T) - \mathrm{tr}(BB^TZZ^T) - \mathrm{tr}(ZZ^TBB^T) + \mathrm{tr}(ZZ^TBB^TZZ^T)$$

cyclic prop of trace

$$\boxed{\mathrm{tr}(BB^T) - \mathrm{tr}(Z^TBB^TZ)} - \cancel{\mathrm{tr}(Z^TBB^TZ)} + \cancel{\mathrm{tr}(\underbrace{Z^TZ}_{I}Z^TBB^TZ)}$$

Formal Analysis

Let $Z_* \in \mathbb{R}^{n\times k}$ contain the top left singular vectors of A – i.e. $Z_* = \arg\min \|A - ZZ^TA\|_F^2$. Similarly, $\bar{\mathbf{Z}} = \arg\min \|\mathbf{C} - ZZ^T\mathbf{C}\|_F^2$.

Claim 1: For any orthonormal $Z \in \mathbb{R}^{n\times k}$, and any matrix B,

$$\|B - ZZ^TB\|_F^2 = \text{tr}(BB^T) - \text{tr}(Z^TBB^TZ).$$

Claim 2: If $\|AA^T - \mathbf{C}\mathbf{C}^T\|_F \leq \frac{\epsilon}{\sqrt{k}}\|A\|_F^2$, then for any orthonormal $Z \in \mathbb{R}^{n\times k}$, $|\text{tr}(Z^T(AA^T - CC^T)Z)| \leq \epsilon\|A\|_F^2$.

$\text{tr}(Z^TAA^TZ) - \text{tr}(Z^TCC^TZ)$

Formal Analysis

Let $Z_* \in \mathbb{R}^{n\times k}$ contain the top left singular vectors of A – i.e. $Z_* = \arg\min \|A - ZZ^T A\|_F^2$. Similarly, $\bar{\mathbf{Z}} = \arg\min \|\mathbf{C} - ZZ^T\mathbf{C}\|_F^2$.

Claim 1: For any orthonormal $Z \in \mathbb{R}^{n\times k}$, and any matrix B,

$$\|B - ZZ^T B\|_F^2 = \operatorname{tr}(BB^T) - \operatorname{tr}(Z^T BB^T Z).$$

Claim 2: If $\|AA^T - \mathbf{C}\mathbf{C}^T\|_F \leq \frac{\epsilon}{\sqrt{k}}\|A\|_F^2$, then for any orthonormal $Z \in \mathbb{R}^{n\times k}$, $\operatorname{tr}(Z^T(AA^T - CC^T)Z) \leq \epsilon\|A\|_F^2$.

Proof from claims:

Formal Analysis

Let $Z_* \in \mathbb{R}^{n\times k}$ contain the top left singular vectors of A – i.e. $Z_* = \arg\min \|A - ZZ^T A\|_F^2$. Similarly, $\bar{\mathbf{Z}} = \arg\min \|\mathbf{C} - ZZ^T\mathbf{C}\|_F^2$.

Claim 1: For any orthonormal $Z \in \mathbb{R}^{n\times k}$, and any matrix B,

$$\|B - ZZ^T B\|_F^2 = \operatorname{tr}(BB^T) - \operatorname{tr}(Z^T BB^T Z).$$

Claim 2: If $\|AA^T - \mathbf{C}\mathbf{C}^T\|_F \leq \frac{\epsilon}{\sqrt{k}}\|A\|_F^2$, then for any orthonormal $Z \in \mathbb{R}^{n\times k}$, $\operatorname{tr}(Z^T(AA^T - CC^T)Z) \leq \epsilon\|A\|_F^2$.

Proof from claims:

top eigenvectors of CC^T

holds by Courant-Fischer.

opt for A

opt C

$$\|\mathbf{C} - \bar{\mathbf{Z}}\bar{\mathbf{Z}}^T\mathbf{C}\|_F^2 \leq \|\mathbf{C} - Z_* Z_*^T\mathbf{C}\|_F^2 \implies \underline{\operatorname{tr}(\bar{\mathbf{Z}}^T\mathbf{C}\mathbf{C}^T\bar{\mathbf{Z}})} \geq \operatorname{tr}(Z_*^T\mathbf{C}\mathbf{C}^T Z_*)$$

$\underline{tr(CC^T)} - tr(Z^T CC^T Z)$

Formal Analysis

Let $Z_* \in \mathbb{R}^{n \times k}$ contain the top left singular vectors of A – i.e. $Z_* = \arg\min \|A - ZZ^T A\|_F^2$. Similarly, $\bar{\mathbf{Z}} = \arg\min \|\mathbf{C} - ZZ^T \mathbf{C}\|_F^2$.

Claim 1: For any orthonormal $Z \in \mathbb{R}^{n \times k}$, and any matrix B,

$$\|B - ZZ^T B\|_F^2 = \mathrm{tr}(BB^T) - \mathrm{tr}(Z^T BB^T Z).$$

Claim 2: If $\|AA^T - \mathbf{C}\mathbf{C}^T\|_F \leq \frac{\epsilon}{\sqrt{k}}\|A\|_F^2$, then for any orthonormal $Z \in \mathbb{R}^{n \times k}$, $\mathrm{tr}(Z^T(AA^T - CC^T)Z) \leq \epsilon\|A\|_F^2$.

$\mathrm{tr}(\bar{Z}^T A A^T \bar{Z}) \pm \epsilon\|A\|_F^2$

$\mathrm{tr}(Z_* A A^T Z_*) \pm \epsilon\|A\|_F^2$

Proof from claims:

$$\|\mathbf{C} - \bar{\mathbf{Z}}\bar{\mathbf{Z}}^T \mathbf{C}\|_F^2 \leq \|\mathbf{C} - Z_* Z_*^T \mathbf{C}\|_F^2 \implies \mathrm{tr}(\bar{\mathbf{Z}}^T \mathbf{C}\mathbf{C}^T \bar{\mathbf{Z}}) \geq \mathrm{tr}(Z_*^T \mathbf{C}\mathbf{C}^T Z_*)$$

$$\implies \mathrm{tr}(\bar{\mathbf{Z}}^T AA^T \bar{\mathbf{Z}}) \geq \mathrm{tr}(Z_*^T AA^T Z_*) - 2\epsilon\|A\|_F^2$$

Formal Analysis

Let $Z_* \in \mathbb{R}^{n\times k}$ contain the top left singular vectors of A – i.e. $Z_* = \arg\min \|A - ZZ^TA\|_F^2$. Similarly, $\bar{\mathbf{Z}} = \arg\min \|\mathbf{C} - ZZ^T\mathbf{C}\|_F^2$.

Claim 1: For any orthonormal $Z \in \mathbb{R}^{n\times k}$, and any matrix B,

$$\|B - ZZ^TB\|_F^2 = \mathrm{tr}(BB^T) - \mathrm{tr}(Z^TBB^TZ).$$

Claim 2: If $\|AA^T - \mathbf{C}\mathbf{C}^T\|_F \leq \frac{\epsilon}{\sqrt{k}}\|A\|_F^2$, then for any orthonormal $Z \in \mathbb{R}^{n\times k}$, $\mathrm{tr}(Z^T(AA^T - CC^T)Z) \leq \epsilon\|A\|_F^2$.

Proof from claims:

$$\|\mathbf{C} - \bar{\mathbf{Z}}\bar{\mathbf{Z}}^T\mathbf{C}\|_F^2 \leq \|\mathbf{C} - Z_*Z_*^T\mathbf{C}\|_F^2 \implies \mathrm{tr}(\bar{\mathbf{Z}}^T\mathbf{C}\mathbf{C}^T\bar{\mathbf{Z}}) \geq \mathrm{tr}(Z_*^T\mathbf{C}\mathbf{C}^TZ_*)$$

$$\implies \mathrm{tr}(\bar{\mathbf{Z}}^TAA^T\bar{\mathbf{Z}}) \geq \mathrm{tr}(Z_*^TAA^TZ_*) - 2\epsilon\|A\|_F^2$$

$$\implies \|A - \bar{\mathbf{Z}}\bar{\mathbf{Z}}^TA\|_F^2 \leq \|A - Z_*Z_*^TA\|_F^2 + 2\epsilon\|A\|_F^2.$$

$\mathrm{tr}(AA^T) - \mathrm{tr}(\bar{Z}^TAA\bar{Z}^T)$

Formal Analysis

Claim 2: If $\|AA^T - CC^T\|_F \leq \frac{\epsilon}{\sqrt{k}}\|A\|_F^2$, then for any orthonormal $Z \in \mathbb{R}^{n\times k}$, $\text{tr}(Z^T(AA^T - CC^T)Z) \leq \epsilon\|A\|_F^2$.

Formal Analysis

Claim 2: If $\|AA^T - CC^T\|_F \leq \frac{\epsilon}{\sqrt{k}}\|A\|_F^2$, then for any orthonormal $Z \in \mathbb{R}^{n\times k}$, $\mathrm{tr}(Z^T(A\underbrace{A^T - CC^T)Z) \leq \epsilon\|A\|}\|_F^2$.

Suffices to show that for any symmetric $B \in \mathbb{R}^{n\times n}$, and any orthonormal $Z \in \mathbb{R}^{n\times k}$, $\mathrm{tr}(Z^TBZ) \leq \sqrt{k}\cdot\|B\|_F$.

$B = AA^T - CC^T$

Formal Analysis

Claim 2: If $\|AA^T - \mathbf{C}\mathbf{C}^T\|_F \leq \frac{\epsilon}{\sqrt{k}}\|A\|_F^2$, then for any orthonormal $Z \in \mathbb{R}^{n \times k}$, $\text{tr}(Z^T(AA^T - \mathbf{C}\mathbf{C}^T)Z) \leq \epsilon\|A\|_F^2$.

Suffices to show that for any symmetric $B \in \mathbb{R}^{n \times n}$, and any orthonormal $Z \in \mathbb{R}^{n \times k}$, $\text{tr}(Z^TBZ) \leq \sqrt{k} \cdot \|B\|_F$.

$$\text{tr}(Z^TBZ) = \sum_{i=1}^{k} z_i^T B z_i$$

$$\begin{bmatrix} z_1^T \\ \vdots \\ z_k^T \end{bmatrix} \begin{bmatrix} B \end{bmatrix} \begin{bmatrix} z_1 \cdots z_k \end{bmatrix}$$

Claim 2: If $\|AA^T - \mathbf{C}\mathbf{C}^T\|_F \leq \frac{\epsilon}{\sqrt{k}}\|A\|_F^2$, then for any orthonormal $Z \in \mathbb{R}^{n\times k}$, $\text{tr}(Z^T(AA^T - \mathbf{C}\mathbf{C}^T)Z) \leq \epsilon\|A\|_F^2$.

Suffices to show that for any symmetric $B \in \mathbb{R}^{n\times n}$, and any orthonormal $Z \in \mathbb{R}^{n\times k}$, $\text{tr}(Z^TBZ) \leq \sqrt{k}\cdot\|B\|_F$.

z_i ith eigenvector of B

$z_i^T B z_i = \lambda_i(B)$

$$\text{tr}(Z^TBZ) = \sum_{i=1}^{k} z_i^T B z_i$$

$$\leq \sum_{i=1}^{k} \lambda_i(B) \qquad \text{(By Courant-Fischer theorem)}$$

Formal Analysis

Claim 2: If $\|AA^T - \mathbf{C}\mathbf{C}^T\|_F \le \frac{\epsilon}{\sqrt{k}}\|A\|_F^2$, then for any orthonormal $Z \in \mathbb{R}^{n\times k}$, $\operatorname{tr}(Z^T(AA^T - \mathbf{C}\mathbf{C}^T)Z) \le \epsilon\|A\|_F^2$.

Suffices to show that for any symmetric $B \in \mathbb{R}^{n\times n}$, and any orthonormal $Z \in \mathbb{R}^{n\times k}$, $\operatorname{tr}(Z^TBZ) \le \sqrt{k}\cdot\|B\|_F$.

$$\begin{aligned}\operatorname{tr}(Z^TBZ) &= \sum_{i=1}^k z_i^TBz_i \\ &\le \sum_{i=1}^k \underline{\lambda_i(B)} \qquad \text{(By Courant-Fischer theorem)} \\ &\le \sqrt{k}\cdot\sqrt{\sum_{i=1}^k \lambda_i(B)^2}\end{aligned}$$

ℓ_1 norm

ℓ_2 norm

Formal Analysis

Claim 2: If $\|AA^T - \mathbf{C}\mathbf{C}^T\|_F \leq \frac{\epsilon}{\sqrt{k}}\|A\|_F^2$, then for any orthonormal $Z \in \mathbb{R}^{n\times k}$, $\mathrm{tr}(Z^T(AA^T - \mathbf{C}\mathbf{C}^T)Z) \leq \epsilon\|A\|_F^2$.

Suffices to show that for any symmetric $B \in \mathbb{R}^{n\times n}$, and any orthonormal $Z \in \mathbb{R}^{n\times k}$, $\mathrm{tr}(Z^TBZ) \leq \sqrt{k}\cdot\|B\|_F$.

$$
\begin{aligned}
\mathrm{tr}(Z^TBZ) &= \sum_{i=1}^{k} z_i^T B z_i \\
&\leq \sum_{i=1}^{k} \lambda_i(B) \qquad \text{(By Courant-Fischer theorem)} \\
&\leq \sqrt{k}\cdot\sqrt{\sum_{i=1}^{k}\lambda_i(B)^2} \leq \sqrt{k}\cdot\sqrt{\sum_{i=1}^{n}\lambda_i(B)^2} .
\end{aligned}
$$

$$\|B\|_F^2 = \sum B_{ij}^2 = \sum \lambda_i(B)^2$$

Formal Analysis

Claim 2: If $\|AA^T - \mathbf{C}\mathbf{C}^T\|_F \leq \frac{\epsilon}{\sqrt{k}}\|A\|_F^2$, then for any orthonormal $Z \in \mathbb{R}^{n\times k}$, $\mathrm{tr}(Z^T(AA^T - \mathbf{C}\mathbf{C}^T)Z) \leq \epsilon\|A\|_F^2$.

Suffices to show that for any symmetric $B \in \mathbb{R}^{n\times n}$, and any orthonormal $Z \in \mathbb{R}^{n\times k}$, $\mathrm{tr}(Z^TBZ) \leq \sqrt{k}\cdot\|B\|_F$.

$$\lambda = \begin{bmatrix} \lambda_1(B) \\ \vdots \\ \lambda_k(B) \end{bmatrix} \qquad \mathbb{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$\mathrm{tr}(Z^TBZ) = \sum_{i=1}^k z_i^T B z_i$$

$$\langle \mathbb{1}, \lambda \rangle \leq \sum_{i=1}^k \lambda_i(B) \quad \text{(By Courant-Fischer theorem)}$$

$$\leq \|\mathbb{1}\|_2 \cdot \|\lambda\|_2 \qquad \sqrt{k} \cdot \sqrt{\textstyle\sum \lambda_i(B)^2}$$

Cauchy-Schwarz.

$$\leq \sqrt{k}\cdot\sqrt{\sum_{i=1}^k \lambda_i(B)^2} \leq \sqrt{k}\cdot\sqrt{\sum_{i=1}^n \lambda_i(B)^2} = \sqrt{k}\cdot\|B\|_F.$$

More Advanced Techniques

$\frac{1}{\varepsilon^2}$

Norm based sampling gives an additive error approximation, $\|A - \bar{\mathbf{Z}}\bar{\mathbf{Z}}^T A\|_F^2 \leq \min_{Z:Z^TZ=I} \|A - ZZ^TA\|_F^2 + 2\epsilon\|A\|_F^2$.

More Advanced Techniques

Norm based sampling gives an additive error approximation,
$\|A - \overline{ZZ}^T A\|_F^2 \leq \min_{Z:Z^TZ=I} \|A - ZZ^TA\|_F^2 + 2\epsilon\|A\|_F^2$.

- Ideally, we would like a relative error approximation,
  $\|A - \overline{ZZ}^T A\|_F^2 \leq (1+\epsilon) \cdot \min_{Z:Z^TZ=I} \|A - ZZ^TA\|_F^2$.

More Advanced Techniques

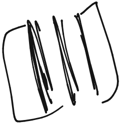
Norm based sampling gives an additive error approximation,
$\|A - \overline{ZZ}^T A\|_F^2 \leq \min_{Z:Z^TZ=I} \|A - ZZ^TA\|_F^2 + 2\epsilon\|A\|_F^2$.

- Ideally, we would like a relative error approximation,
$\|A - \overline{ZZ}^T A\|_F^2 \leq (1+\epsilon) \cdot \min_{Z:Z^TZ=I} \|A - ZZ^TA\|_F^2$.
- This can be achieved with more advanced non-uniform sampling techniques, based on leverage scores or adaptive sampling.
- Also possible using Johnson-Lindenstrauss type random projection.

Adaptive Sampling

Given an input matrix $A \in \mathbb{R}^{n \times d}$ and rank parameter $k \ll \min(n, d)$.

- Initialize probabilities $p_i = \cancel{1/n}\ \frac{\|A_{:,i}\|_2^2}{\|A\|_F^2}$ for $i \in [n]$.
- Initialize list of columns $C = \{\}$ and orthonormal matrix $V = 0$.
- For $j = 1, 2, \ldots t$
  - Set a column $c_j \in \{A_{:,1}, \ldots, A_{:,n}\}$ to $A_{:,i}$ with probability p_i and add c_j to C.
  - Let $V \in \mathbb{R}^{n \times j}$ have orthonormal columns spanning the columns in C.
  - For all $i \in [n]$, let $p_i = \frac{\|A_{:,i} - VV^T A_{:,i}\|_2^2}{\|A - VV^T A\|_F^2}$.
- Return the top k left singular values of $AV \in \mathbb{R}^{n \times t}$.



Adaptive Sampling

A $\quad \text{rank}(A) = 3 \quad k = 2$

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

$\text{rank}(ZZ^TA) = 2$

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ & & 0 & 0 & 0 \end{bmatrix}$$

$$Z = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\begin{array}{ccccc} p_1 = 1/3 & p_2 = 1/3 & p_3 = 1/3 & p_4 = p_5 = 0 \end{array}$$

$$A - ZZ^TA = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

$$\|A_{:,i}\|_2^2 - \underbrace{\|ZZ^TA_{:,i}\|_2^2}_{=\ \|Z^TA_{:,i}\|_2^2}$$

naked