COMPSCI 690RA: Randomized Algorithms and Probabilistic Data Analysis

Prof. Cameron Musco

University of Massachusetts Amherst. Spring 2022.

Lecture 6

Logistics (Lots of Them)

- 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:

- 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.

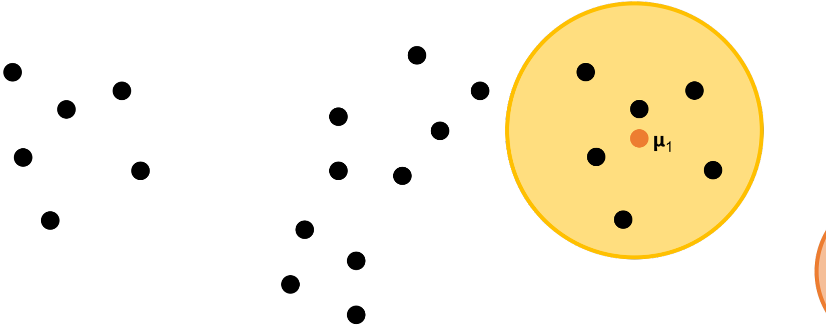
Today: More applications of non-uniform and adaptive sampling to clustering and low-rank approximation.

- 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.



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:

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.

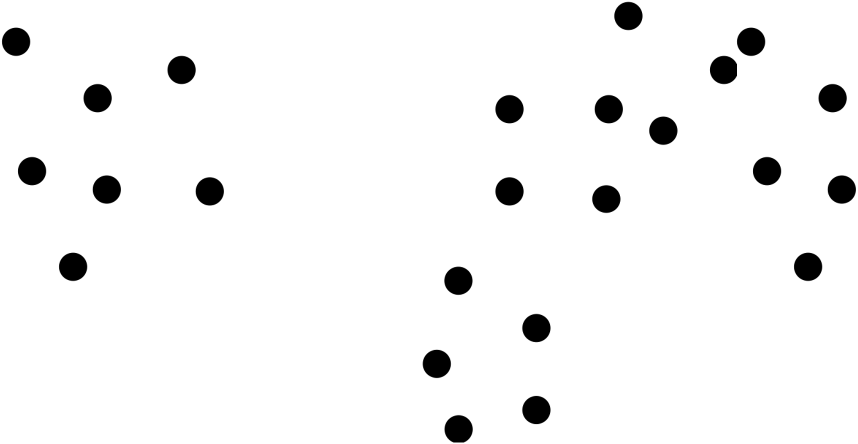
Lloyd’s Algorithm

k-means++

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

- 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 $C = \{c_1, \ldots, c_k\}$.

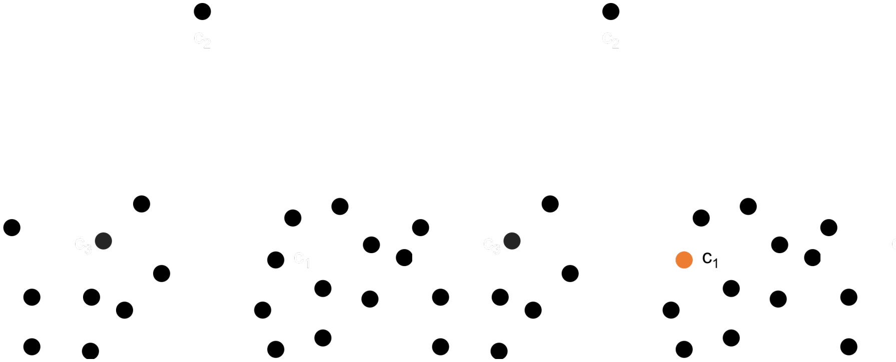
k-means++



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? This deterministic variant can be foiled by outliers.



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

k-means++ Analysis

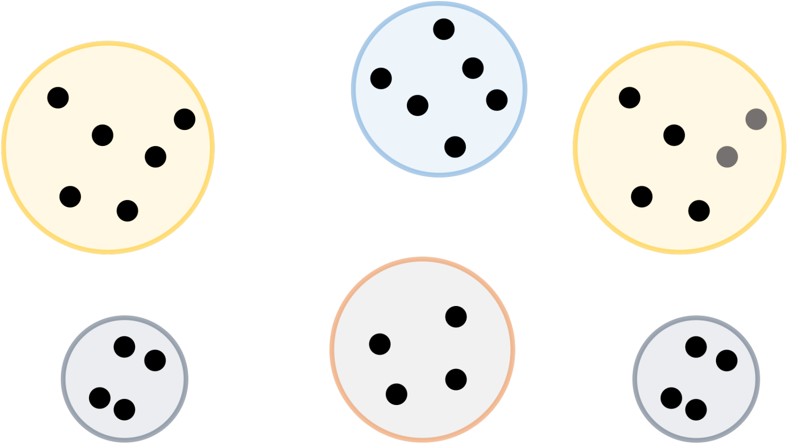
Proof Outline:

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-means++ Analysis

Proof Outline:

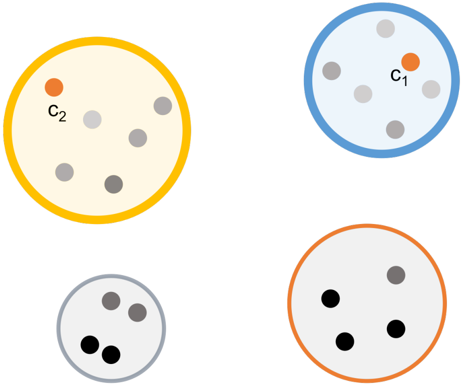
1. Let $C_1, \ldots, C_k$ the clusters corresponding to centers $c_1, \ldots, c_k$ and $\mu(C_1), \ldots, \mu(C_k)$ be their centroids. Let $A_1, \ldots, A_k$ be the optimal clusters. We will show:

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

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.

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 $\mathbb{E}[\phi(\mathcal{X}_c)] \lesssim \phi_{OPT}(\mathcal{X}_c)$

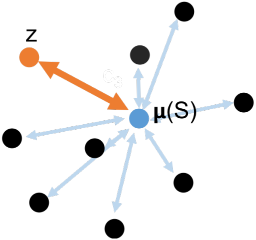
k-means++ Analysis

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.

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

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



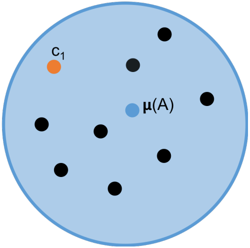
Proof: $\sum_{x \in S} \|x - z\|_2^2 = \sum_{x \in S} \|(x - \mu(S)) + (\mu(S) - z)\|_2^2 =$
$\sum_{x \in S} \|x - \mu(S)\|_2^2 + \sum_{x \in S} \|\mu(S) - z\|_2^2 + \sum_{x \in S} 2\langle x - \mu(S), \mu(S) - z\rangle$

First Cluster Bound

Lemma

Let A be some cluster in the optimal cluster set $A_1, \dots, A_k$. 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 $\phi_{OPT}(A) = \sum_{x \in A} \|x - \mu(A)\|_2^2$.

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



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

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

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

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

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.

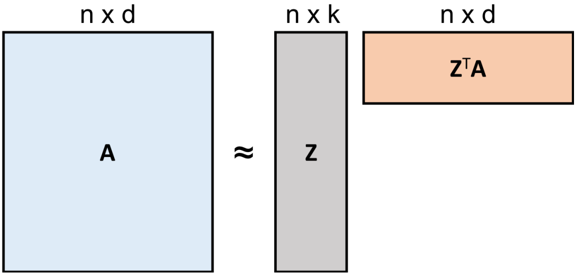
Randomized Low-Rank approximation

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:

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

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



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

Sampling Based Algorithm

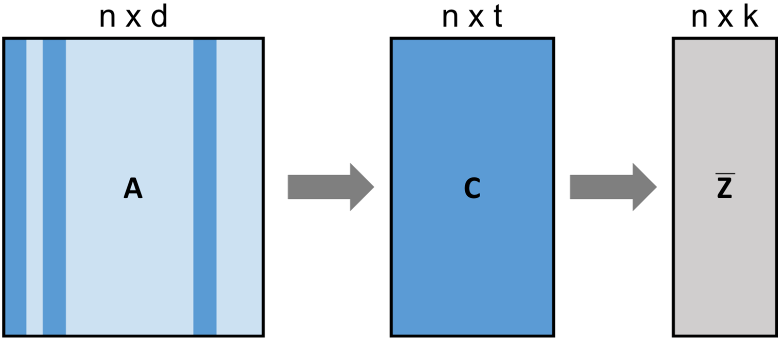
We will analysis a simple non-uniform sampling based algorithm for low-rank approximation, that gives a near optimal solution in $O(nd + nk^2)$ time.

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



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 $\overline{\mathbf{Z}} \in \mathbb{R}^{n \times k}$ satisfying with probability at least $1 - \delta$:

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

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

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

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 $\overline{\mathbf{Z}}$ 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. $Z_* = \arg\min \|A - ZZ^T A\|_F^2$. 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,

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

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

Proof from claims:

$$\begin{aligned}
\|\mathbf{C} - \bar{\mathbf{Z}}\bar{\mathbf{Z}}^T\mathbf{C}\|_F^2 \leq \|\mathbf{C} - Z_* Z_*^T \mathbf{C}\|_F^2 &\implies \text{tr}(\bar{\mathbf{Z}}^T\mathbf{C}\mathbf{C}^T\bar{\mathbf{Z}}) \geq \text{tr}(Z_*^T\mathbf{C}\mathbf{C}^T Z_*) \\
&\implies \text{tr}(\bar{\mathbf{Z}}^T AA^T\bar{\mathbf{Z}}) \geq \text{tr}(Z_*^T AA^T Z_*) - 2\epsilon\|A\|_F^2 \\
&\implies \|A - \bar{\mathbf{Z}}\bar{\mathbf{Z}}^T A\|_F^2 \leq \|A - Z_* Z_*^T A\|_F^2 + 2\epsilon\|A\|_F^2.
\end{aligned}$$

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 $Z \in \mathbb{R}^{n\times k}$, $\text{tr}(Z^TBZ) \leq \sqrt{k}\cdot\|B\|_F$.

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

More Advanced Techniques

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

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