DIMENSIONALITY REDUCTION FOR K-MEANS AND LOW RANK APPROXIMATION

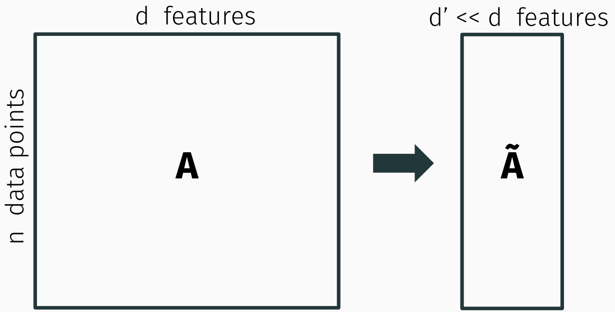
Michael Cohen, Sam Elder, Cameron Musco,
Christopher Musco, Mădălina Persu

Massachusetts Institute of Technology
(currently at Yahoo Labs, NYC)

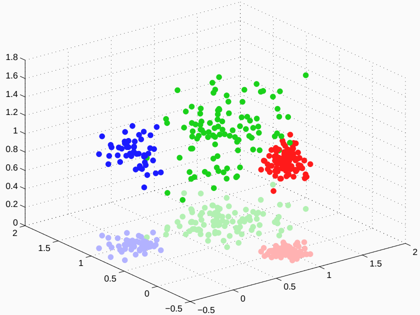
Simple techniques to accelerate algorithms for:

- k-means clustering
- principal component analysis (PCA)
- **constrained low rank approximation**

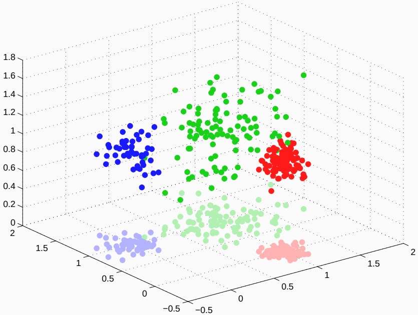
Replace large, high dimensional dataset with low dimensional sketch.



Solution on sketch $\tilde{\mathbf{A}}$ should approximate original solution.

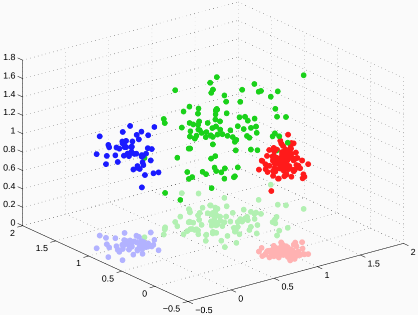


Solution on sketch $\tilde{\mathbf{A}}$ should approximate original solution.



Dimensionality reduction algorithm is ideally fast, memory efficient – often randomization is used.

Solution on sketch $\tilde{\mathbf{A}}$ should approximate original solution.



Simultaneously improves runtime, memory requirements, communication cost, etc.

WELL STUDIED APPROACH

Standard paradigm for "randomized numerical linear algebra".

Standard paradigm for "randomized numerical linear algebra".

- Obtaining pre-conditioners for matrix inversion
- Constrained regression (i.e. non-negative least squares)
- Fast SVMs, kernel approximation, algebraic graph theory, etc.

WELL STUDIED APPROACH

Standard paradigm for "randomized numerical linear algebra".

- Obtaining pre-conditioners for matrix inversion
- Constrained regression (i.e. non-negative least squares)
- Fast SVMs, kernel approximation, algebraic graph theory, etc.
- Low rank approximation, principal component analysis

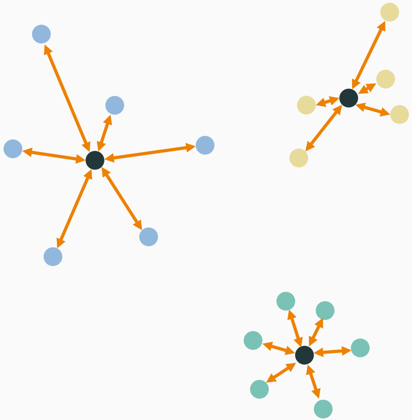
- Extremely common objective function for clustering

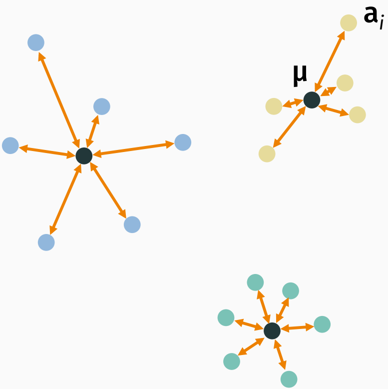
- Extremely common objective function for clustering
- Choose k clusters to minimize total intra-cluster variance

- Extremely common objective function for clustering
- Choose k clusters to minimize total intra-cluster variance

- Extremely common objective function for clustering
- Choose k clusters to minimize total intra-cluster variance

- Extremely common objective function for clustering
- Choose k clusters to minimize total intra-cluster variance





$$\min_{C} Cost(\mathbf{A}, C) = \sum_{i=1}^{n} \|\mathbf{a}_i - \boldsymbol{\mu}\left(C[\mathbf{a}_i]\right)\|_2^2$$

$$\min_{C} Cost(\mathsf{A}, C) = \sum_{i=1}^{n} \left\| \mathsf{a}_i - \boldsymbol{\mu}\left(C[\mathsf{a}_i]\right) \right\|_2^2$$

$$\min_{C} Cost(\mathbf{A}, C) = \sum_{i=1}^{n} \|\mathbf{a}_i - \boldsymbol{\mu}\left(C[\mathbf{a}_i]\right)\|_2^2$$

- NP-hard, even for fixed dimension d or fixed k.

$$\min_{C} Cost(\mathsf{A}, C) = \sum_{i=1}^{n} \|\mathsf{a}_i - \boldsymbol{\mu}\left(C[\mathsf{a}_i]\right)\|_2^2$$

- NP-hard, even for fixed dimension d or fixed k.
- Several $(1+\epsilon)$ and constant factor approximation algorithms.

$$\min_{C} Cost(\mathbf{A}, C) = \sum_{i=1}^{n} \|\mathbf{a}_i - \boldsymbol{\mu}\left(C[\mathbf{a}_i]\right)\|_2^2$$

- NP-hard, even for fixed dimension d or fixed k.
- Several $(1+\epsilon)$ and constant factor approximation algorithms.
- In practice: Lloyd's heuristic (i.e. "the k-means algorithm") with k-means++ initialization is used. $O(\log k)$ approximation guaranteed, typically performs much better.

$$\min_{C} Cost(\mathbf{A}, C) = \sum_{i=1}^{n} \|\mathbf{a}_i - \boldsymbol{\mu}\left(C[\mathbf{a}_i]\right)\|_2^2$$

- NP-hard, even for fixed dimension d or fixed k.
- Several $(1+\epsilon)$ and constant factor approximation algorithms.
- In practice: Lloyd's heuristic (i.e. "the k-means algorithm") with k-means++ initialization is used. $O(\log k)$ approximation guaranteed, typically performs much better.

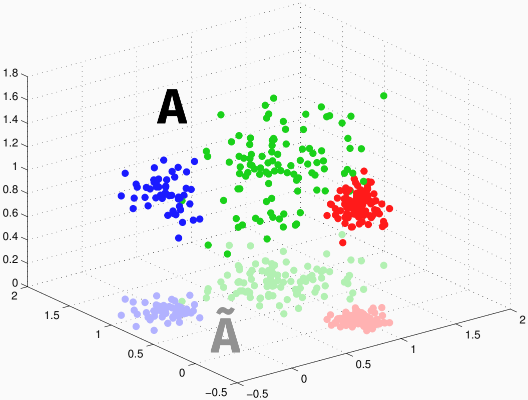
Dimensionality reduction can speed up any of these algorithms.

$$\min_{C} Cost(\mathbf{A}, C) = \sum_{i=1}^{n} \|\mathbf{a}_i - \boldsymbol{\mu}(C[\mathbf{a}_i])\|_2^2$$

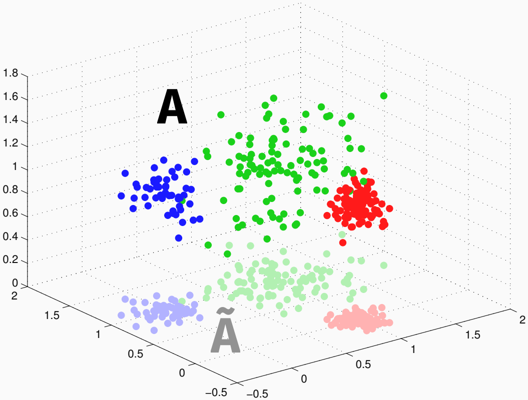
- NP-hard, even for fixed dimension d or fixed k.
- Several $(1+\epsilon)$ and constant factor approximation algorithms.
- In practice: Lloyd's heuristic (i.e. "the k-means algorithm") with k-means++ initialization is used. $O(\log k)$ approximation guaranteed, typically performs much better.

Expecially powerful for Lloyd's algorithm – most of your time is spent computing distances between points!

Let me convince you *something* is possible.

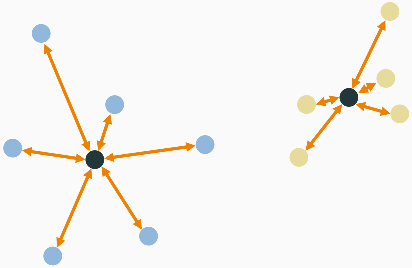


If $Cost(\tilde{\mathsf{A}}, C) \approx Cost(\mathsf{A}, C)$ for all C,
$\min_C Cost(\tilde{\mathsf{A}}, C) \approx \min_C Cost(\mathsf{A}, C)$.

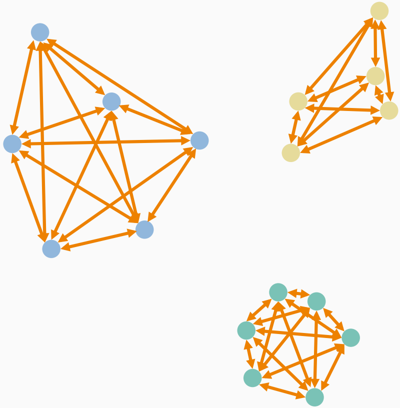


Objective: $Cost(\tilde{A}, C) \approx Cost(A, C)$

Can rewrite cost function:



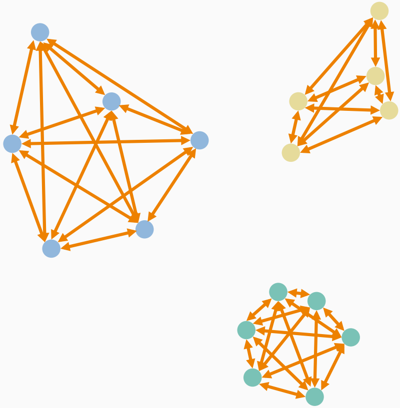
Can rewrite cost function:



$$\sum_{i=1}^{n} \|\mathbf{a}_i - \boldsymbol{\mu}(C[\mathbf{a}_i])\|_2^2 = \sum_{l=1}^{k} \frac{1}{2|C_j|} \sum_{i,j \in C_j} \|\mathbf{a}_i - \mathbf{a}_j\|_2^2$$

OBJECTIVE FUNCTION IN TERMS OF DISTANCES

Can rewrite cost function:



$$\sum_{i=1}^{n} \|\mathbf{a}_i - \boldsymbol{\mu}(C[\mathbf{a}_i])\|_2^2 = \sum_{l=1}^{k} \frac{1}{2|C_j|} \sum_{i,j \in C_j} \|\mathbf{a}_i - \mathbf{a}_j\|_2^2$$

One option: preserve $\|\mathbf{a}_i - \mathbf{a}_j\|_2^2$ for all i, j:

One option: preserve $\|\mathbf{a}_i - \mathbf{a}_j\|_2^2$ for all i, j:

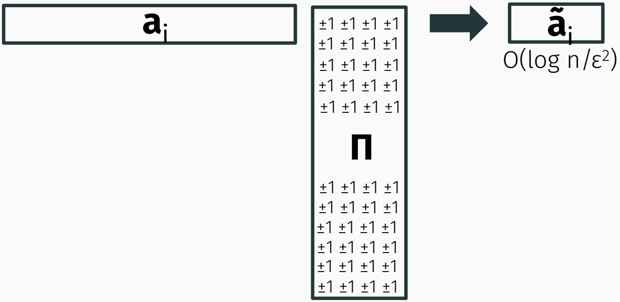
$$\|\tilde{\mathbf{a}}_i - \tilde{\mathbf{a}}_j\|_2^2 = (1 \pm \epsilon)\|\mathbf{a}_i - \mathbf{a}_j\|_2^2 \rightarrow$$

One option: preserve $\|\mathbf{a}_i - \mathbf{a}_j\|_2^2$ for all i, j:

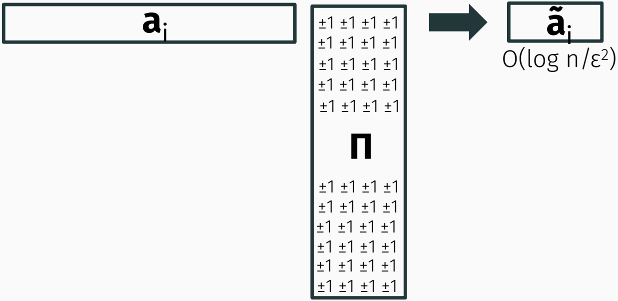
$$\|\tilde{\mathbf{a}}_i - \tilde{\mathbf{a}}_j\|_2^2 = (1 \pm \epsilon)\|\mathbf{a}_i - \mathbf{a}_j\|_2^2 \rightarrow$$

$$Cost(\tilde{\mathbf{A}}, C) = (1 \pm \epsilon) Cost(\mathbf{A}, C)$$

If we have *n* points:



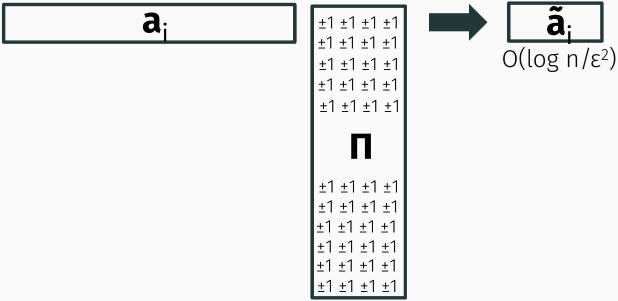
If we have *n* points:



$$\|\tilde{\mathsf{a}}_i - \tilde{\mathsf{a}}_j\|_2^2 = (1 \pm \epsilon)\|\mathsf{a}_i - \mathsf{a}_j\|_2^2$$

JOHNSON-LINDENSTRAUSS LEMMA

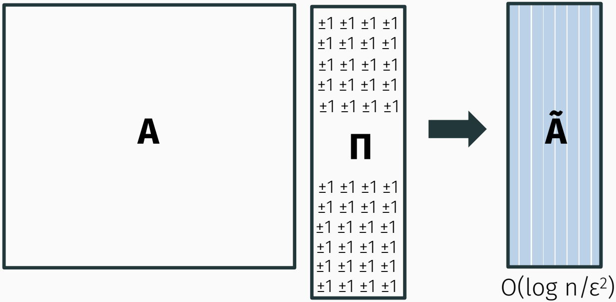
If we have n points:



$$\|\tilde{\mathbf{a}}_i - \tilde{\mathbf{a}}_j\|_2^2 = (1 \pm \epsilon)\|\mathbf{a}_i - \mathbf{a}_j\|_2^2$$

Roughly equivalent to projecting points to a random $O(\log n/\epsilon^2)$ dimensional subspace.

$$\min_C Cost(\tilde{A}, C) \leq (1+\epsilon) \min_C Cost(A, C)$$



Pros:

Pros:

- Simple and fast implementation

Pros:

- Simple and fast implementation
- Easily adaptable to parallel, distributed, and streaming environments

Pros:

- Simple and fast implementation
- Easily adaptable to parallel, distributed, and streaming environments

Cons:

Pros:

- Simple and fast implementation
- Easily adaptable to parallel, distributed, and streaming environments

Cons:

- $O(\log n/\epsilon^2)$ dimension scales with problem size (number of points)

Pros:

- Simple and fast implementation
- Easily adaptable to parallel, distributed, and streaming environments

Cons:

- $O(\log n/\epsilon^2)$ dimension scales with problem size (number of points)
- ϵ^2 dependence and constant factor on $O()$ can be costly

Pros:

- Simple and fast implementation
- Easily adaptable to parallel, distributed, and streaming environments

Cons:

- $O(\log n/\epsilon^2)$ dimension scales with problem size (number of points)
- ϵ^2 dependence and constant factor on $O()$ can be costly
- Problem specific analysis – doesn't generalize

Reframe as a linear algebra problem. Results:

ALTERNATIVE APPROACH

Reframe as a linear algebra problem. Results:

- Wider variety of algorithms. Several beat Johnson-Lindenstrauss random projection (in theory and practice)

Reframe as a linear algebra problem. Results:

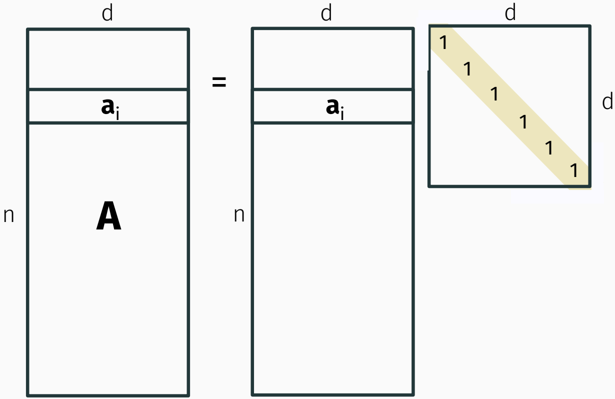
- Wider variety of algorithms. Several beat Johnson-Lindenstrauss random projection (in theory and practice)
- Analysis extends to many additional problems

This approach has led to lots of papers:

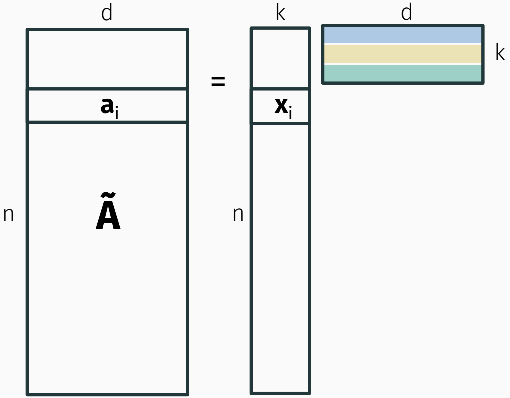
- Drineas, Frieze, Kannan, Vempala, Vinay '04
- Boutsidis, Magdon-Ismail '13
- Feldman, Schmidt, Sohler '13
- Boutsidis, Zouzias, Mahoney, Drineas '09 '10 '15

Review:

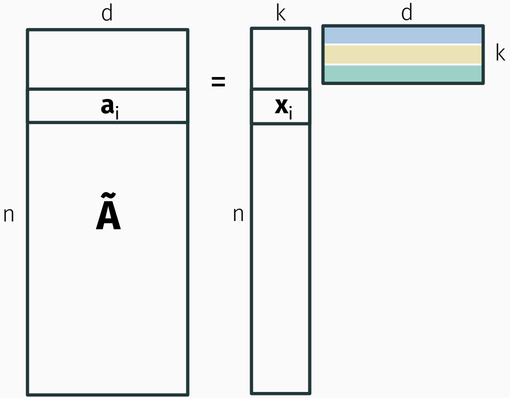
Review:



Review:

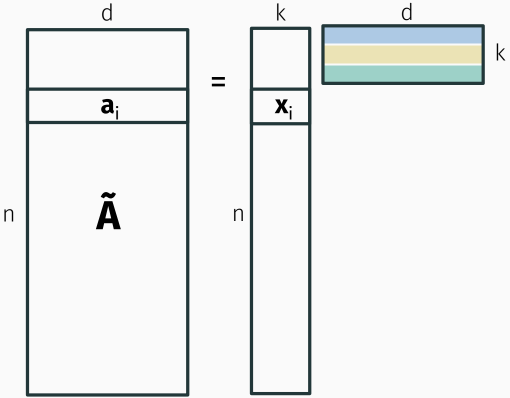


Review:



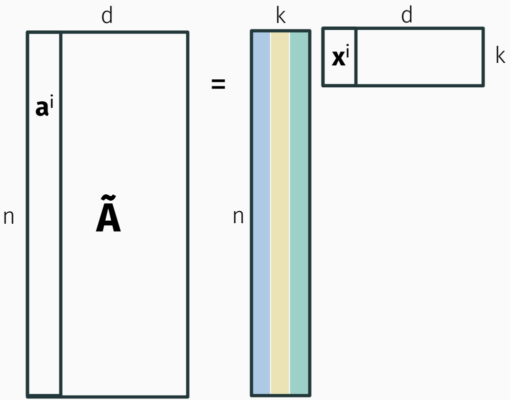
Want $\|\mathbf{A} - \tilde{\mathbf{A}}\|$ to be small.

Review:



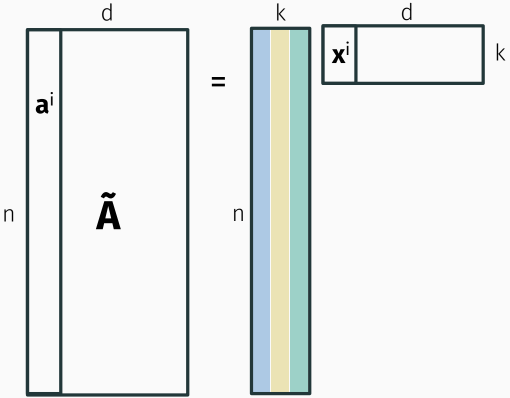
Want $\|\mathbf{A} - \tilde{\mathbf{A}}\|_F^2$ to be small.

Review:



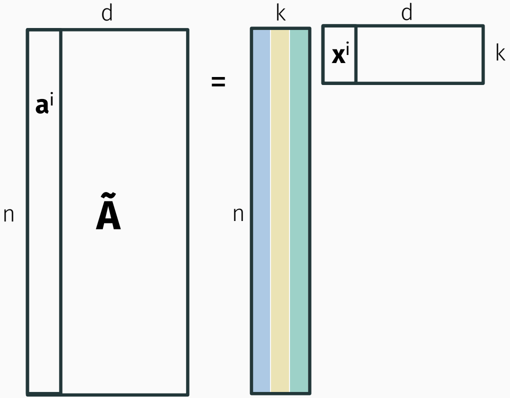
Want $\|A - \tilde{A}\|_F^2$ to be small.

Review:



Want $\|A - \tilde{A}\|_F^2$ to be small.

Review:

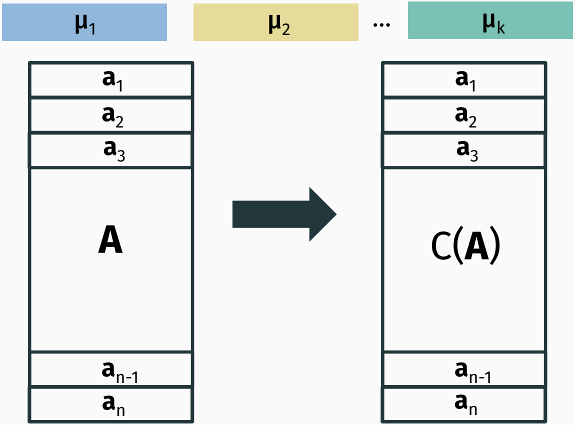


Given set of columns $\mathbf{C}$, best approximation is $\tilde{\mathbf{A}} = \text{proj}_{\mathbf{C}}(\mathbf{A})$.

k-means clustering == low rank approximation

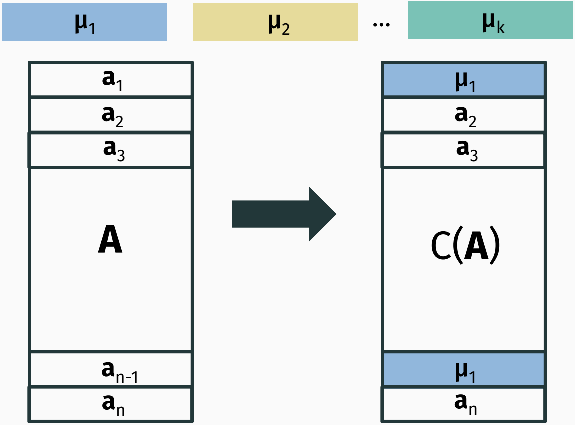
k-means clustering == low rank approximation

$$\min \sum_{i=1}^{n} \|\mathrm{a}_i - \boldsymbol{\mu}(\mathrm{a}_i)\|_2^2$$



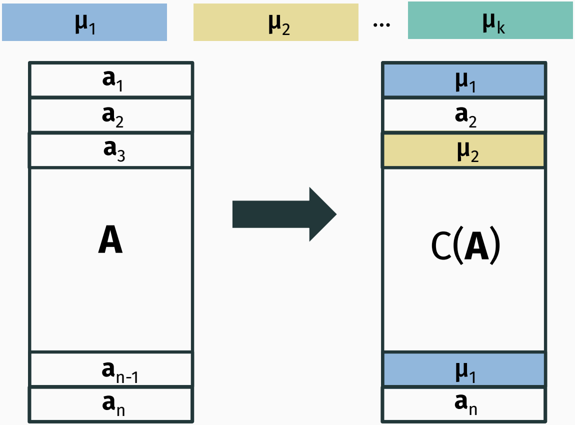
k-means clustering == low rank approximation

$$\min \sum_{i=1}^{n} \|\mathbf{a}_i - \boldsymbol{\mu}(\mathbf{a}_i)\|_2^2$$



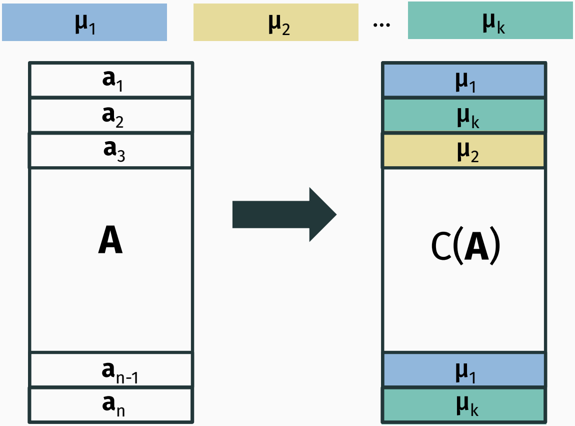
k-means clustering == low rank approximation

$$\min \sum_{i=1}^{n} \|\mathrm{a}_i - \boldsymbol{\mu}(\mathrm{a}_i)\|_2^2$$



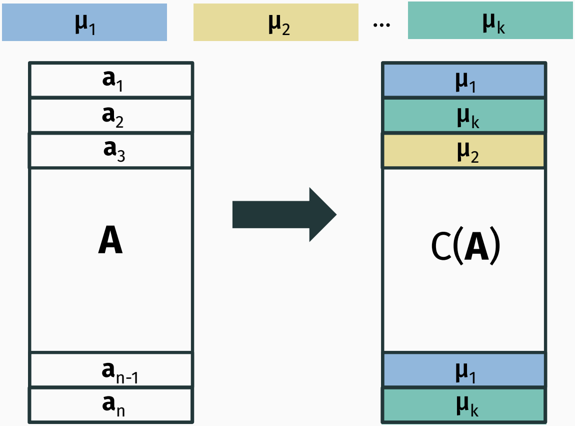
k-means clustering == low rank approximation

$$\min \sum_{i=1}^{n} \|a_i - \mu(a_i)\|_2^2$$



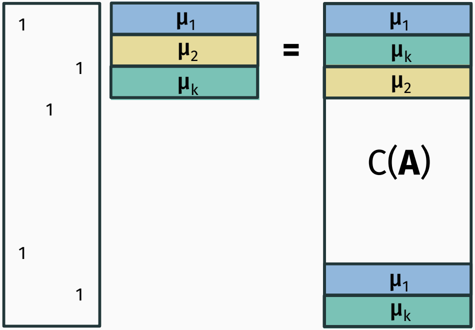
k-means clustering == low rank approximation

$$\min \sum_{i=1}^{n} \|a_i - \boldsymbol{\mu}(a_i)\|_2^2 = \|A - C(A)\|_F^2$$

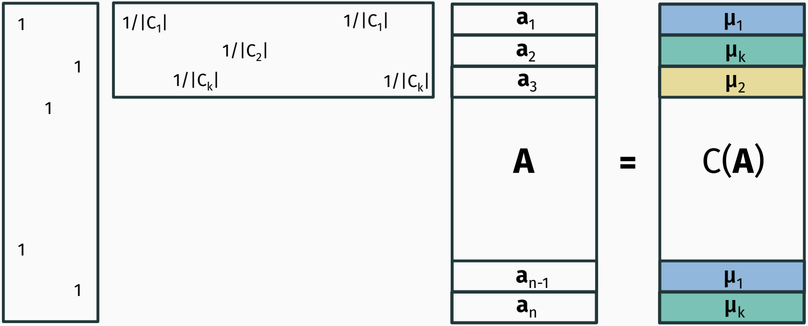


$C(\mathbf{A})$ is actually a projection of $\mathbf{A}$'s columns onto a rank k subspace [Boutsidis, Drineas, Mahoney, Zouzias '11]

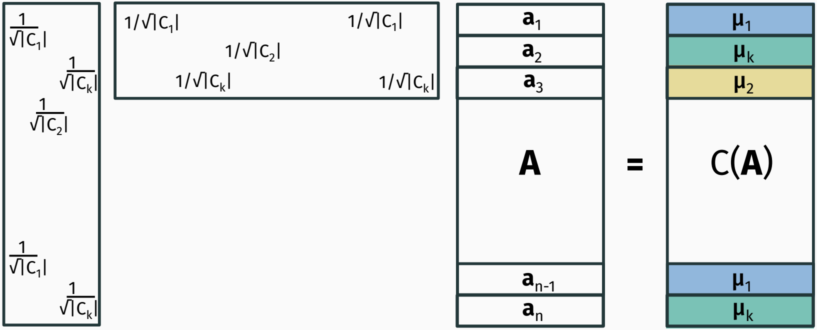
$C(\mathbf{A})$ is actually a projection of $\mathbf{A}$'s columns onto a rank k subspace [Boutsidis, Drineas, Mahoney, Zouzias '11]



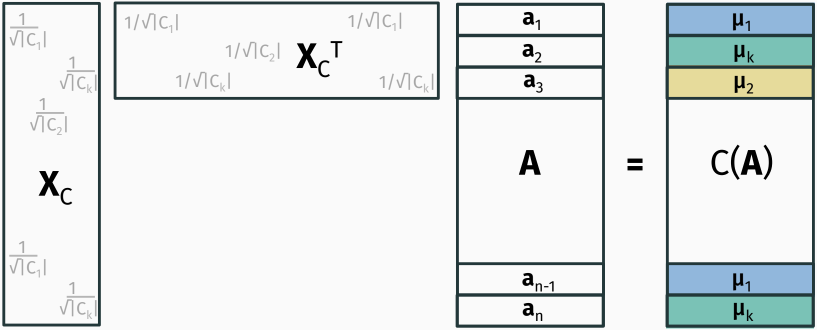
C(A) is actually a projection of A's columns onto a rank k subspace [Boutsidis, Drineas, Mahoney, Zouzias '11]



C(A) is actually a projection of A's columns onto a rank k subspace [Boutsidis, Drineas, Mahoney, Zouzias '11]



$C(A)$ is actually a projection of A's columns onto a rank k subspace [Boutsidis, Drineas, Mahoney, Zouzias '11]



$$\min_{C} \sum_{i=1}^{n} \|\mathbf{a}_i - \boldsymbol{\mu}(C[\mathbf{a}_i])\|_2^2 \Longrightarrow \min_{rank(\mathbf{X})=k, \mathbf{X}\in\mathcal{S}} \|\mathbf{A} - \mathbf{X}\mathbf{X}^\top\mathbf{A}\|_F^2$$

Where $\mathbf{X}$ is a rank k orthonormal matrix and for k-means $\mathcal{S}$ is the set of all clustering indicator matrices.

$$\min_{C} \sum_{i=1}^{n} \|\mathbf{a}_i - \boldsymbol{\mu}\left(C[\mathbf{a}_i]\right)\|_2^2 \implies \min_{rank(\mathbf{X})=k, \mathbf{X}\in\mathcal{S}} \|\mathbf{A} - \mathbf{X}\mathbf{X}^\top\mathbf{A}\|_F^2$$

Where $\mathbf{X}$ is a rank k orthonormal matrix and for k-means $\mathcal{S}$ is the set of all clustering indicator matrices.

- General form for constrained low rank approximation.

$$\min_{C} \sum_{i=1}^{n} \|\mathbf{a}_i - \boldsymbol{\mu}(C[\mathbf{a}_i])\|_2^2 \implies \min_{rank(\mathbf{X})=k, \mathbf{X}\in\mathcal{S}} \|\mathbf{A} - \mathbf{X}\mathbf{X}^\top \mathbf{A}\|_F^2$$

Where $\mathbf{X}$ is a rank k orthonormal matrix and for k-means $\mathcal{S}$ is the set of all clustering indicator matrices.

- General form for constrained low rank approximation.
- Set $\mathcal{S} = \{\text{All rank } k \text{ orthonormal matrices}\}$ for principal component analysis (unconstrained low rank approx.)

SO WHAT?

Want to solve any problem like $\min_{rank(\mathbf{X})=k, \mathbf{X}\in\mathcal{S}} \|\mathbf{A} - \mathbf{X}\mathbf{X}^\top\mathbf{A}\|_F^2$.

SO WHAT?

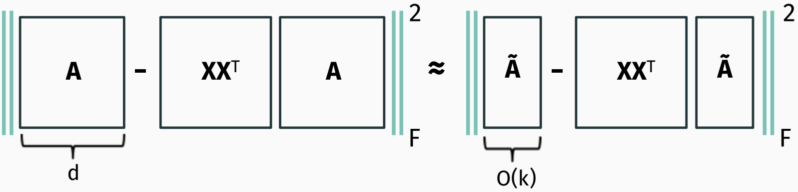
Want to solve any problem like $\min_{rank(\mathbf{X})=k, \mathbf{X}\in\mathcal{S}} \|\mathbf{A} - \mathbf{X}\mathbf{X}^\top\mathbf{A}\|_F^2$.

For all rank k $\mathbf{X}$, $\|\tilde{\mathbf{A}} - \mathbf{X}\mathbf{X}^\top\tilde{\mathbf{A}}\|_F^2 \approx \|\mathbf{A} - \mathbf{X}\mathbf{X}^\top\mathbf{A}\|_F^2$

SO WHAT?

Want to solve any problem like $\min_{rank(\mathbf{X})=k, \mathbf{X} \in \mathcal{S}} \|\mathbf{A} - \mathbf{X}\mathbf{X}^\top \mathbf{A}\|_F^2$.

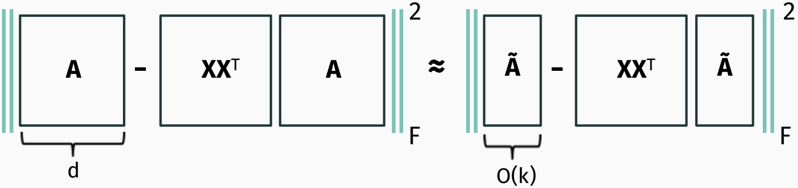
$$\text{For all rank } k \ \mathbf{X}, \ \|\tilde{\mathbf{A}} - \mathbf{X}\mathbf{X}^\top \tilde{\mathbf{A}}\|_F^2 \approx \|\mathbf{A} - \mathbf{X}\mathbf{X}^\top \mathbf{A}\|_F^2$$



SO WHAT?

Want to solve any problem like $\min_{rank(X)=k, X\in\mathcal{S}} \|A - XX^\top A\|_F^2$.

$$\text{For all rank } k \text{ } X, \ \|\tilde{A} - XX^\top \tilde{A}\|_F^2 \approx \|A - XX^\top A\|_F^2$$



Projection-Cost Preserving Sketch

Specifically, we want:

$$\text{for all } \mathbf{X},\ \ \|\tilde{\mathbf{A}} - \mathbf{X}\mathbf{X}^\top \tilde{\mathbf{A}}\|_F^2 \approx \|\mathbf{A} - \mathbf{X}\mathbf{X}^\top \mathbf{A}\|_F^2$$

Specifically, we want:

$$\text{for all } \mathbf{X},\ \|\tilde{\mathbf{A}} - \mathbf{X}\mathbf{X}^\top\tilde{\mathbf{A}}\|_F^2 = (1 \pm \epsilon)\|\mathbf{A} - \mathbf{X}\mathbf{X}^\top\mathbf{A}\|_F^2$$

Specifically, we want:

$$\text{for all } \mathbf{X}, \; \|\tilde{\mathbf{A}} - \mathbf{X}\mathbf{X}^\top\tilde{\mathbf{A}}\|_F^2 + c = (1 \pm \epsilon)\|\mathbf{A} - \mathbf{X}\mathbf{X}^\top\mathbf{A}\|_F^2$$

Specifically, we want:

$$\text{for all } \mathbf{X},\ \|\tilde{\mathbf{A}} - \mathbf{X}\mathbf{X}^\top\tilde{\mathbf{A}}\|_F^2 + c = (1 \pm \epsilon)\|\mathbf{A} - \mathbf{X}\mathbf{X}^\top\mathbf{A}\|_F^2$$

If we find an $\mathbf{X}$ that gives γ approximate solution for $\tilde{\mathbf{A}}$:

$$\|\tilde{\mathbf{A}} - \mathbf{X}\mathbf{X}^\top\tilde{\mathbf{A}}\|_F^2 \leq \gamma\|\tilde{\mathbf{A}} - \widetilde{\mathbf{X}\mathbf{X}}_{opt}^\top\tilde{\mathbf{A}}\|_F^2$$

Specifically, we want:

$$\text{for all } \mathbf{X}, \;\; \|\tilde{\mathbf{A}} - \mathbf{X}\mathbf{X}^\top\tilde{\mathbf{A}}\|_F^2 + c = (1 \pm \epsilon)\|\mathbf{A} - \mathbf{X}\mathbf{X}^\top\mathbf{A}\|_F^2$$

If we find an $\mathbf{X}$ that gives γ approximate solution for $\tilde{\mathbf{A}}$:

$$\|\tilde{\mathbf{A}} - \mathbf{X}\mathbf{X}^\top\tilde{\mathbf{A}}\|_F^2 \leq \gamma\|\tilde{\mathbf{A}} - \widetilde{\mathbf{X}\mathbf{X}}_{opt}^\top\tilde{\mathbf{A}}\|_F^2$$

then:

$$\|\mathbf{A} - \mathbf{X}\mathbf{X}^\top\mathbf{A}\|_F^2 \leq \gamma \cdot (1 + \epsilon)\|\mathbf{A} - \mathbf{X}\mathbf{X}_{opt}^\top\mathbf{A}\|_F^2$$

Specifically, we want:

$$\text{for all } \mathbf{X},\ \|\tilde{\mathbf{A}} - \mathbf{X}\mathbf{X}^\top\tilde{\mathbf{A}}\|_F^2 + c = (1 \pm \epsilon)\|\mathbf{A} - \mathbf{X}\mathbf{X}^\top\mathbf{A}\|_F^2$$

If we find an $\mathbf{X}$ that gives γ approximate solution for $\tilde{\mathbf{A}}$:

$$\|\tilde{\mathbf{A}} - \mathbf{X}\mathbf{X}^\top\tilde{\mathbf{A}}\|_F^2 \leq \gamma\|\tilde{\mathbf{A}} - \widetilde{\mathbf{X}\mathbf{X}}_{opt}^\top\tilde{\mathbf{A}}\|_F^2$$

then:

$$\|\mathbf{A} - \mathbf{X}\mathbf{X}^\top\mathbf{A}\|_F^2 \leq \gamma \cdot (1 + \epsilon)\|\mathbf{A} - \mathbf{X}\mathbf{X}_{opt}^\top\mathbf{A}\|_F^2$$

See coresets of Feldman, Schmidt, Sohler '13.

What we have seen so far:

What we have seen so far:

- k-means clustering is just constrained k rank approximation.

RECAP

What we have seen so far:

- k-means clustering is just constrained k rank approximation.
- Sufficient to construct a projection-cost preserving sketch $\tilde{\mathbf{A}}$ that approximates distance from $\mathbf{A}$ to any rank k subspace.

What we have seen so far:

- k-means clustering is just constrained k rank approximation.
- Sufficient to construct a projection-cost preserving sketch $\tilde{\mathbf{A}}$ that approximates distance from $\mathbf{A}$ to any rank k subspace.
- Stronger guarantee than has been sought in prior work on approximate PCA via sketching

Toolbox of dimensionality reduction algorithms for obtaining projection-cost preserving matrix sketches.

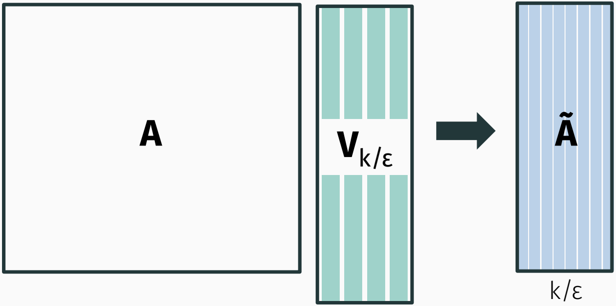
Toolbox of dimensionality reduction algorithms for obtaining projection-cost preserving matrix sketches.

- deterministic & randomized - unified analysis - many applications beyond k-means

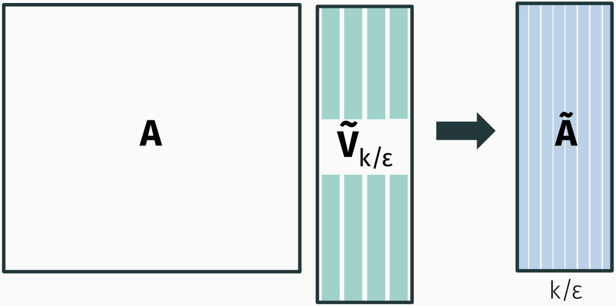
What techniques give $(1+\epsilon)$ projection-cost preservation?

	Previous Work			Our Results	
Technique	Reference	Dimensions	Error	Dimensions	Error
SVD	Feldman, Schmidt, Sohler '13	$O(k/\epsilon^2)$	$1+\epsilon$	$\lceil k/\epsilon \rceil$	$1+\epsilon$
Approximate SVD	Boutsidis, Drineas, Mahoney, Zouzias '11	k	$2+\epsilon$	$\lceil k/\epsilon \rceil$	$1+\epsilon$
Random Projection	"	$O(k/\epsilon^2)$	$2+\epsilon$	$O(k/\epsilon^2)$ $O(\log k/\epsilon^2)$	$1+\epsilon$ $9+\epsilon$
Column Sampling	"	$O(k\log k/\epsilon^2)$	$3+\epsilon$	$O(k\log k/\epsilon^2)$	$1+\epsilon$
Deterministic Column Selection	Boutsidis, Magdon-Ismail '13	$r > k$	$O(n/r)$	$O(k/\epsilon^2)$	$1+\epsilon$
Non-oblivious Projection	NA	NA	NA	$O(k/\epsilon)$	$1+\epsilon$

	Previous Work			Our Results	
Technique	Reference	Dimensions	Error	Dimensions	Error
SVD	Feldman, Schmidt, Sohler '13	$O(k/\epsilon^2)$	$1+\epsilon$	$\lceil k/\epsilon \rceil$	$1+\epsilon$



	Previous Work			Our Results	
Technique	Reference	Dimensions	Error	Dimensions	Error
Approximate SVD	Boutsidis, Drineas, Mahoney, Zouzias '11	k	$2+\epsilon$	$\lceil k/\epsilon \rceil$	$1+\epsilon$

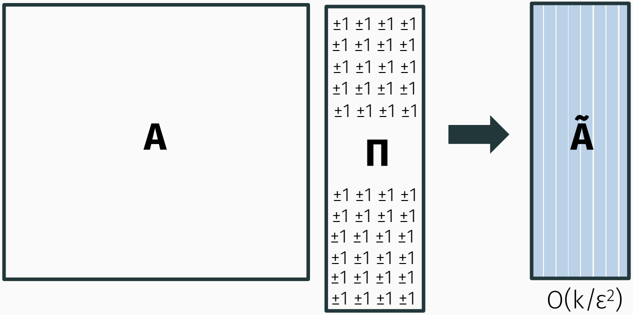


	Previous Work			Our Results	
Technique	Reference	Dimensions	Error	Dimensions	Error
Approximate SVD	Boutsidis, Drineas, Mahoney, Zouzias '11	k	$2+\epsilon$	$\lceil k/\epsilon \rceil$	$1+\epsilon$

- Practically *very* useful for k-means
- No constant factors on k/ϵ, and typically many fewer dimensions are required (Kappmeier, Schmidt, Schmidt '15)

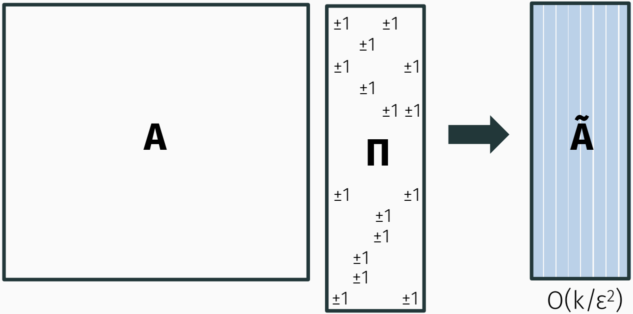
JOHNSON-LINDENSTRAUSS RANDOM PROJECTION

	Previous Work			Our Results	
Technique	Reference	Dimensions	Error	Dimensions	Error
Random Projection	Boutsidis, Drineas, Mahoney, Zouzias '11	$O(k/\epsilon^2)$	$2+\epsilon$	$O(k/\epsilon^2)$	$1+\epsilon$
		$O(\log n/\epsilon^2)$	$1+\epsilon$	$O(\log k/\epsilon^2)$	$9+\epsilon$

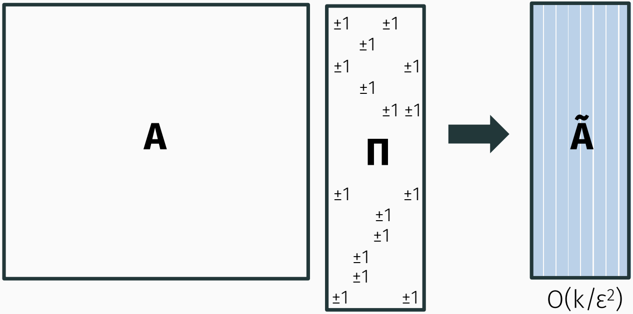


JOHNSON-LINDENSTRAUSS RANDOM PROJECTION

	Previous Work			Our Results	
Technique	Reference	Dimensions	Error	Dimensions	Error
Random Projection	Boutsidis, Drineas, Mahoney, Zouzias '11	$O(k/\epsilon^2)$	$2+\epsilon$	$O(k/\epsilon^2)$	$1+\epsilon$
		$O(\log n/\epsilon^2)$	$1+\epsilon$	$O(\log k/\epsilon^2)$	$9+\epsilon$



	Previous Work			Our Results	
Technique	Reference	Dimensions	Error	Dimensions	Error
Random Projection	Boutsidis, Drineas, Mahoney, Zouzias '11	$O(k/\epsilon^2)$	$2+\epsilon$	$O(k/\epsilon^2)$	$1+\epsilon$
		$O(\log n/\epsilon^2)$	$1+\epsilon$	$O(\log k/\epsilon^2)$	$9+\epsilon$



	Previous Work			Our Results	
Technique	Reference	Dimensions	Error	Dimensions	Error
Random Projection	Boutsidis, Drineas, Mahoney, Zouzias '11	$O(k/\epsilon^2)$	$2+\epsilon$	$O(k/\epsilon^2)$	$1+\epsilon$
		$O(\log n/\epsilon^2)$	$1+\epsilon$	$O(\log k/\epsilon^2)$	$9+\epsilon$

- First sketch with dimension sublinear in k. Can $(9+\epsilon)$ be improved?

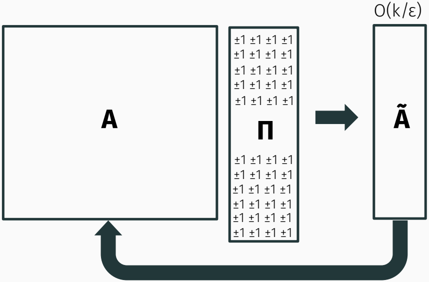
JOHNSON-LINDENSTRAUSS RANDOM PROJECTION

	Previous Work			Our Results	
Technique	Reference	Dimensions	Error	Dimensions	Error
Random Projection	Boutsidis, Drineas, Mahoney, Zouzias '11	$O(k/\epsilon^2)$	$2+\epsilon$	$O(k/\epsilon^2)$	$1+\epsilon$
		$O(\log n/\epsilon^2)$	$1+\epsilon$	$O(\log k/\epsilon^2)$	$9+\epsilon$

- First sketch with dimension sublinear in k. Can $(9+\epsilon)$ be improved?
- Sketch is data oblivious
    - Lowest communication distributed k-means (improves on Balcan, Kanchanapally, Liang, Woodruff '14)
    - Streaming principal component analysis in a single pass

Standard sketches for low rank approximation (Sarlós '06, Clarkson, Woodruff '13, etc):

Standard sketches for low rank approximation (Sarlós ‘06, Clarkson, Woodruff ‘13, etc):

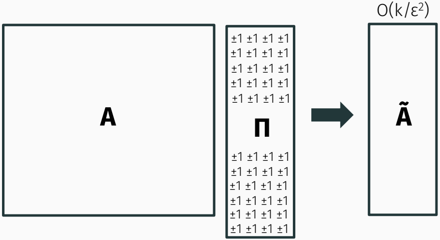


$$\|\mathbf{A} - (\mathbf{P}_{\tilde{\mathbf{A}}}\mathbf{A})_k\|_F^2 \leq (1+\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2.$$

span($\tilde{\mathbf{A}}$) contains a good low rank approximation for **A**, but we must return to **A** to find it.

Projection-cost preserving sketch for low rank approximation:

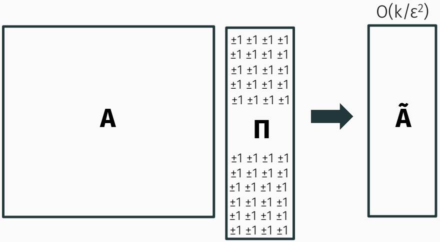
Projection-cost preserving sketch for low rank approximation:



$$\|\mathbf{A} - \mathbf{V}_k \mathbf{V}_k^\top \mathbf{A}\|_F^2 \leq (1+\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2.$$

Additional ϵ dependence for stronger sketch.

Projection-cost preserving sketch for low rank approximation:

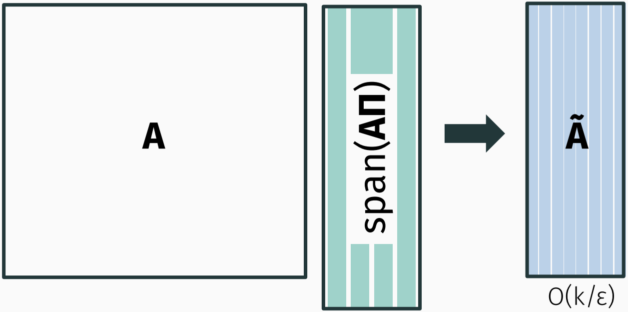


$$\|\mathbf{A} - \mathbf{V}_k \mathbf{V}_k^\top \mathbf{A}\|_F^2 \leq (1+\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2.$$

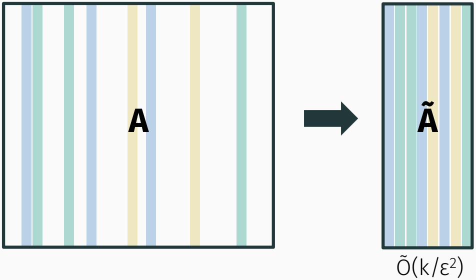
Additional ϵ dependence for stronger sketch. Is it required for oblivious approximate PCA? (See Ghashami, Liberty, Phillips, Woodruff ‘14/15)

TWO SHOT PROJECTION

	Previous Work			Our Results	
Technique	Reference	Dimensions	Error	Dimensions	Error
Non-oblivious Randomized Projection	Sarlós '06	NA	NA	$O(k/\epsilon)$	$1+\epsilon$



	Previous Work			Our Results	
Technique	**Reference**	**Dimensions**	**Error**	**Dimensions**	**Error**
Column Sampling	Boutsidis, Drineas, Mahoney, Zouzias '11	$O(k \log k/\epsilon^2)$	$3+\epsilon$	$O(k \log k/\epsilon^2)$	$1+\epsilon$
Deterministic Column Selection	Boutsidis, Magdon-Ismail '13	$r > k$	$O(n/r)$	$O(k/\epsilon^2)$	$1+\epsilon$



	Previous Work			Our Results	
Technique	Reference	Dimensions	Error	Dimensions	Error
Column Sampling	Boutsidis, Drineas, Mahoney, Zouzias '11	$O(k \log k/\epsilon^2)$	$3+\epsilon$	$O(k \log k/\epsilon^2)$	$1+\epsilon$

- **A** not only contains a small set of columns that span a good low rank approximation, but a small (reweighted) set whose top principal components approximate those of **A**.

	Previous Work			Our Results	
Technique	Reference	Dimensions	Error	Dimensions	Error
Column Sampling	Boutsidis, Drineas, Mahoney, Zouzias '11	$O(k \log k/\epsilon^2)$	$3+\epsilon$	$O(k \log k/\epsilon^2)$	$1+\epsilon$

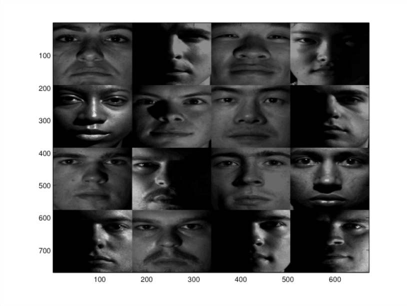
- $\mathbf{A}$ not only contains a small set of columns that span a good low rank approximation, but a small (reweighted) set whose top principal components approximate those of $\mathbf{A}$.
- First single shot sampling based dimensionality reduction for $(1+\epsilon)$ error low rank approximation.

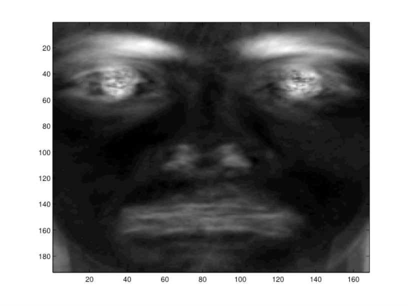
	Previous Work			Our Results	
Technique	Reference	Dimensions	Error	Dimensions	Error
Column Sampling	Boutsidis, Drineas, Mahoney, Zouzias '11	$O(k \log k/\epsilon^2)$	$3+\epsilon$	$O(k \log k/\epsilon^2)$	$1+\epsilon$

- $\mathbf{A}$ not only contains a small set of columns that span a good low rank approximation, but a small (reweighted) set whose top principal components approximate those of $\mathbf{A}$.
- First single shot sampling based dimensionality reduction for $(1+\epsilon)$ error low rank approximation.
    - Work in progress: single-pass streaming column subset selection and iterative sampling algorithms for the SVD.

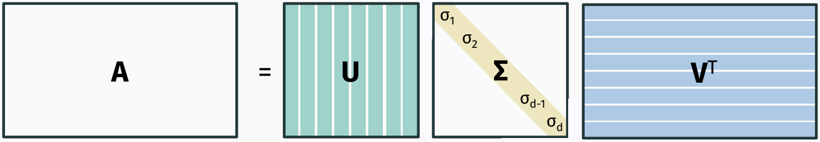
Natural (unsupervised) feature selection metric:

combination of leverage score with respect to top k subspace and residuals of columns after projection to this subspace.

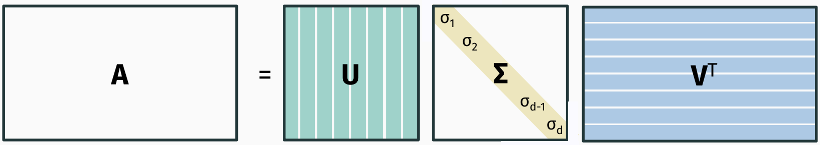




Singular Value Decomposition (SVD):

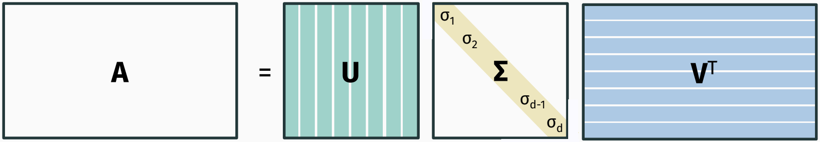


Singular Value Decomposition (SVD):



$\mathbf{U}, \mathbf{V}$ have orthonormal columns. $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d$.

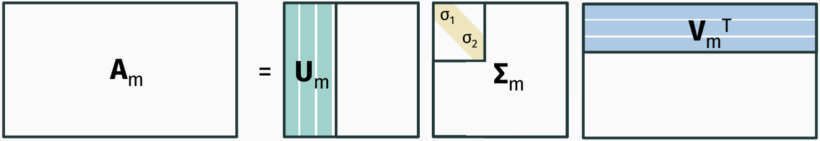
Singular Value Decomposition (SVD):



$\mathbf{U}, \mathbf{V}$ have orthonormal columns. $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d$.

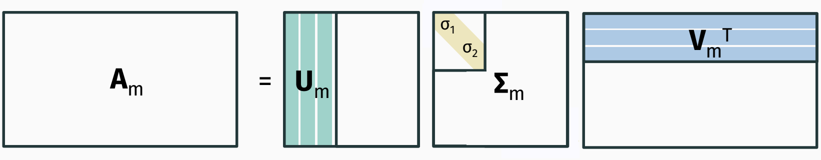
Same as Principal Component Analysis (PCA) if we mean-center $\mathbf{A}$'s columns/rows.

Partial Singular Value Decomposition (SVD):



$$\|A - A_m\|_F^2 = \min_{rank(X)=m} \|A - XX^\top A\|_F^2$$

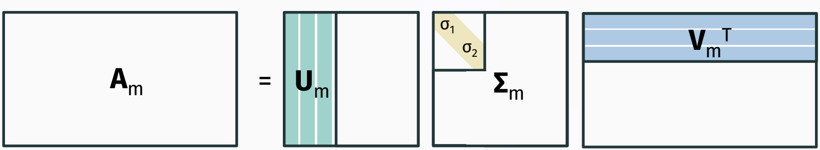
Partial Singular Value Decomposition (SVD):



$$\|A - A_m\|_F^2 = \min_{rank(X)=m} \|A - XX^\top A\|_F^2$$

Claim: $\|A_{k/\epsilon} - XX^\top A_{k/\epsilon}\|_F^2 + c = (1 \pm \epsilon)\|A - XX^\top A\|_F^2$

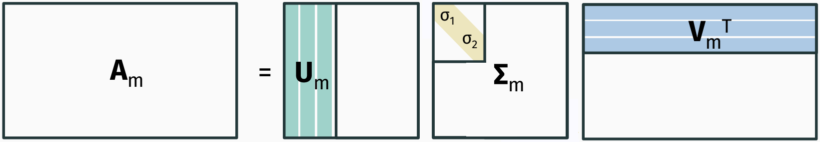
Partial Singular Value Decomposition (SVD):



$$\|\mathbf{A} - \mathbf{A}_m\|_F^2 = \min_{rank(\mathbf{X})=m} \|\mathbf{A} - \mathbf{X}\mathbf{X}^\top \mathbf{A}\|_F^2$$

Claim: $\|\mathbf{U}_{k/\epsilon}\mathbf{\Sigma}_{k/\epsilon} - \mathbf{X}\mathbf{X}^\top \mathbf{U}_{k/\epsilon}\mathbf{\Sigma}_{k/\epsilon}\|_F^2 + c = (1 \pm \epsilon)\|\mathbf{A} - \mathbf{X}\mathbf{X}^\top \mathbf{A}\|_F^2$

Partial Singular Value Decomposition (SVD):



$$\|\mathbf{A} - \mathbf{A}_m\|_F^2 = \min_{rank(\mathbf{X})=m} \|\mathbf{A} - \mathbf{X}\mathbf{X}^\top \mathbf{A}\|_F^2$$

Claim: $\|\mathbf{A}_{k/\epsilon} - \mathbf{X}\mathbf{X}^\top \mathbf{A}_{k/\epsilon}\|_F^2 + c = (1 \pm \epsilon)\|\mathbf{A} - \mathbf{X}\mathbf{X}^\top \mathbf{A}\|_F^2$

We work with $\mathbf{A}_{k/\epsilon}$ for simplicity. Denote $(\mathbf{A} - \mathbf{A}_{k/\epsilon})$ as $\mathbf{A}_{\setminus k/\epsilon}$

Claim: $\|\mathbf{A}_{k/\epsilon} - \mathbf{X}\mathbf{X}^\top \mathbf{A}_{k/\epsilon}\|_F^2 = (1 \pm \epsilon)\|\mathbf{A} - \mathbf{X}\mathbf{X}^\top \mathbf{A}\|_F^2$

Claim: $\|A_{k/\epsilon} - XX^\top A_{k/\epsilon}\|_F^2 = (1 \pm \epsilon)\|A - XX^\top A\|_F^2$

Split into (row) orthogonal pairs:

$$\boxed{\mathbf{A}} = \boxed{\underbrace{\mathbf{A}_{k/\epsilon}}_{\text{head}}} + \boxed{\underbrace{\mathbf{A}_{\setminus k/\epsilon}}_{\text{tail}}}$$

Claim: $\|A_{k/\epsilon} - XX^\top A_{k/\epsilon}\|_F^2 = (1 \pm \epsilon)\|A - XX^\top A\|_F^2$

Split into (row) orthogonal pairs:

$$\boxed{\mathbf{A}} = \boxed{\begin{matrix}\mathbf{A}_{k/\varepsilon} \\ \text{head}\end{matrix}} + \boxed{\begin{matrix}\mathbf{A}_{\setminus k/\varepsilon} \\ \text{tail}\end{matrix}}$$

$\|A - XX^\top A\|_F^2$

Claim: $\|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}_{k/\epsilon}\|_F^2 + c = (1 \pm \epsilon)\|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}\|_F^2$

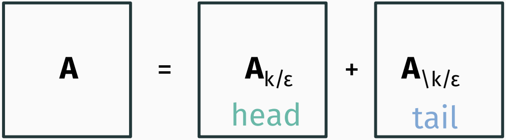
Split into (row) orthogonal pairs:

$$\boxed{\mathbf{A}} = \boxed{\underset{\text{head}}{\mathbf{A}_{k/\varepsilon}}} + \boxed{\underset{\text{tail}}{\mathbf{A}_{\setminus k/\varepsilon}}}$$

$\|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}\|_F^2$

Claim: $\|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}_{k/\epsilon}\|_F^2 + c = (1 \pm \epsilon)\|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}\|_F^2$

Split into (row) orthogonal pairs:



$$= \|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}_{k/\epsilon}\|_F^2 + \|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}_{\setminus k/\epsilon}\|_F^2 = \|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}\|_F^2$$

ANALYSIS FOR SVD

Claim: $\|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}_{k/\epsilon}\|_F^2 + c = (1 \pm \epsilon)\|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}\|_F^2$

Split into (row) orthogonal pairs:

$$\boxed{\mathbf{A}} = \boxed{\begin{matrix}\mathbf{A}_{k/\epsilon} \\ \text{head}\end{matrix}} + \boxed{\begin{matrix}\mathbf{A}_{\setminus k/\epsilon} \\ \text{tail}\end{matrix}}$$

$$= \|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}_{k/\epsilon}\|_F^2 + \|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}_{\setminus k/\epsilon}\|_F^2 = \|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}\|_F^2$$

$$= \|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}_{k/\epsilon}\|_F^2 + \|\mathbf{A}_{\setminus k/\epsilon}\|_F^2 - \|\mathbf{X}\mathbf{X}^\top\mathbf{A}_{\setminus k/\epsilon}\|_F^2 = \|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}\|_F^2$$

Claim: $\|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}_{k/\epsilon}\|_F^2 + c = (1 \pm \epsilon)\|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}\|_F^2$

Split into (row) orthogonal pairs:

$$\boxed{\mathbf{A}} = \boxed{\begin{matrix}\mathbf{A}_{k/\varepsilon} \\ \text{head}\end{matrix}} + \boxed{\begin{matrix}\mathbf{A}_{\setminus k/\varepsilon} \\ \text{tail}\end{matrix}}$$

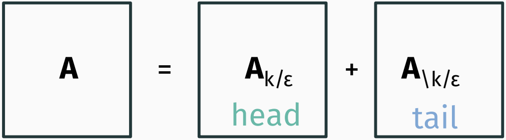
$$= \|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}_{k/\epsilon}\|_F^2 + \|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}_{\setminus k/\epsilon}\|_F^2 = \|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}\|_F^2$$

$$= \|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}_{k/\epsilon}\|_F^2 + c - \|\mathbf{X}\mathbf{X}^\top\mathbf{A}_{\setminus k/\epsilon}\|_F^2 = \|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}\|_F^2$$

ANALYSIS FOR SVD

Claim: $\|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}_{k/\epsilon}\|_F^2 + c = (1 \pm \epsilon)\|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}\|_F^2$

Split into (row) orthogonal pairs:



$$= \|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}_{k/\epsilon}\|_F^2 + \|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}_{\setminus k/\epsilon}\|_F^2 = \|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}\|_F^2$$

$$= \|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}_{k/\epsilon}\|_F^2 + c - \|\mathbf{X}\mathbf{X}^\top\mathbf{A}_{\setminus k/\epsilon}\|_F^2 = \|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}\|_F^2$$

In words, the projection cost for $\mathbf{A}$ is explained by the cost over $\mathbf{A}_{k/\epsilon}$ plus the cost over $\mathbf{A}_{\setminus k/\epsilon}$. We want to argue that ignoring the tail term is fine.

In words, the projection cost for $\mathbf{A}$ is explained by the cost over $\mathbf{A}_{k/\epsilon}$ plus the cost over $\mathbf{A}_{\setminus k/\epsilon}$. We want to argue that ignoring the tail term is fine.

Need to show that:

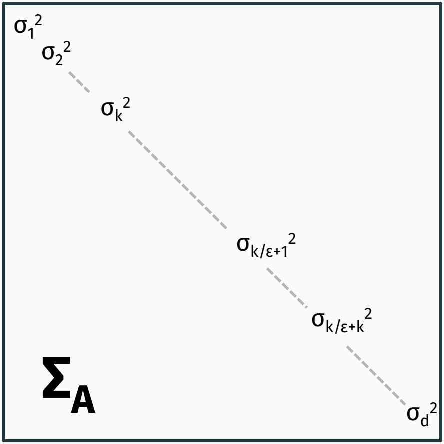
$$\|\mathbf{X}\mathbf{X}^\top \mathbf{A}_{\setminus k/\epsilon}\|_F^2 \leq \epsilon \|(\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}\|_F^2$$

In words, the projection cost for $\mathbf{A}$ is explained by the cost over $\mathbf{A}_{k/\epsilon}$ plus the cost over $\mathbf{A}_{\setminus k/\epsilon}$. We want to argue that ignoring the tail term is fine.

Need to show that:

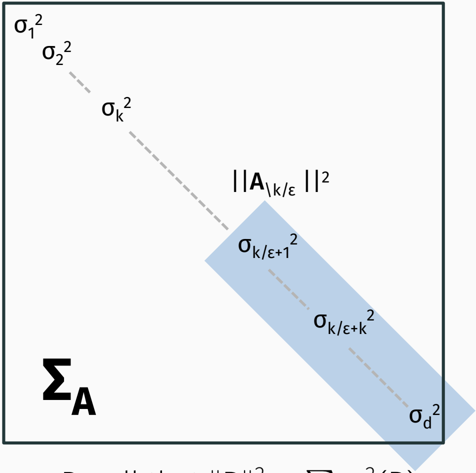
$$\|\mathbf{X}\mathbf{X}^\top\mathbf{A}_{\setminus k/\epsilon}\|_F^2 \leq \epsilon\|\mathbf{A}-\mathbf{A}_k\|_F^2$$

$$\|XX^\top A_{\setminus k/\epsilon}\|_F^2 \leq \epsilon\|A - A_k\|_F^2$$



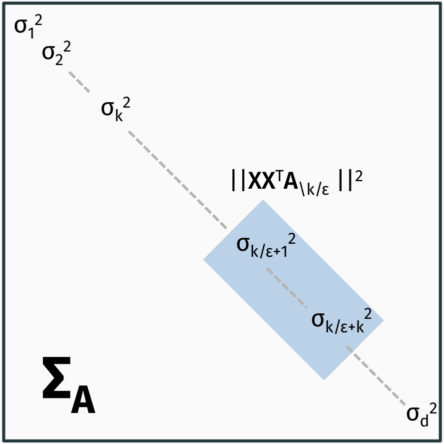
Recall that $\|B\|_F^2 = \sum_i \sigma_i^2(B)$

$$\|\mathbf{X}\mathbf{X}^\top \mathbf{A}_{\setminus k/\epsilon}\|_F^2 \leq \epsilon \|\mathbf{A} - \mathbf{A}_k\|_F^2$$



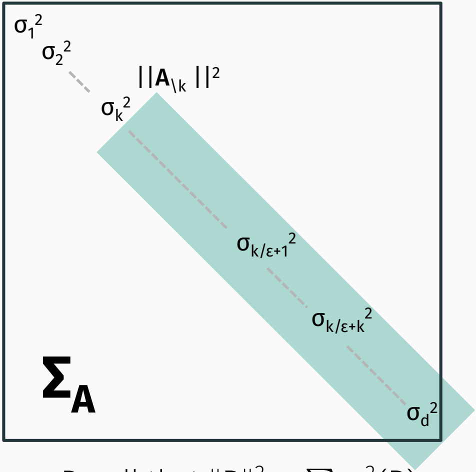
Recall that $\|\mathbf{B}\|_F^2 = \sum_i \sigma_i^2(\mathbf{B})$

$$\|XX^\top A_{\setminus k/\epsilon}\|_F^2 \leq \epsilon \|A - A_k\|_F^2$$



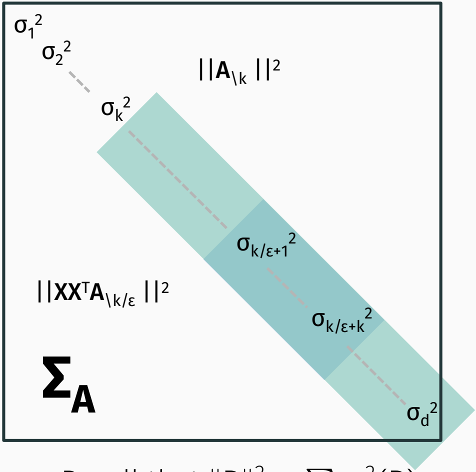
Recall that $\|B\|_F^2 = \sum_i \sigma_i^2(B)$

$$\|\mathbf{X}\mathbf{X}^\top \mathbf{A}_{\setminus k/\epsilon}\|_F^2 \leq \epsilon\|\mathbf{A} - \mathbf{A}_k\|_F^2$$



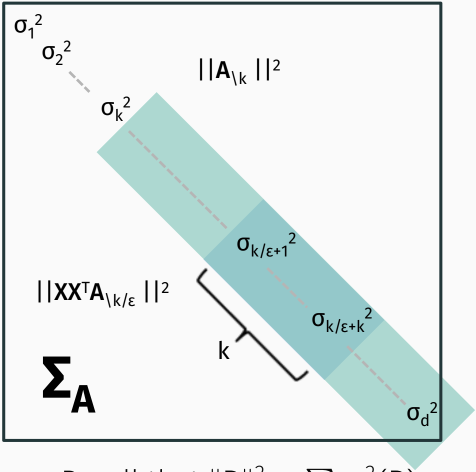
Recall that $\|\mathbf{B}\|_F^2 = \sum_i \sigma_i^2(\mathbf{B})$

$$\|XX^\top A_{\setminus k/\epsilon}\|_F^2 \leq \epsilon \|A - A_k\|_F^2$$



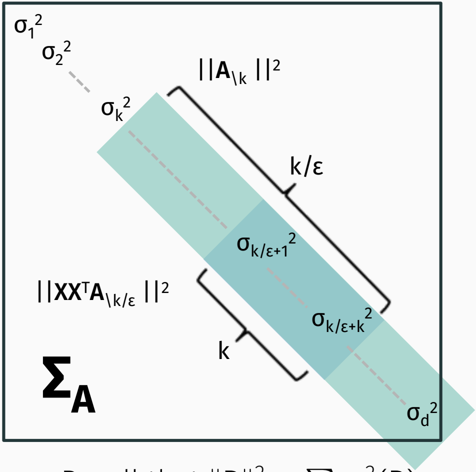
Recall that $\|B\|_F^2 = \sum_i \sigma_i^2(B)$

$$\|XX^\top A_{\setminus k/\epsilon}\|_F^2 \leq \epsilon\|A - A_k\|_F^2$$



Recall that $\|B\|_F^2 = \sum_i \sigma_i^2(B)$

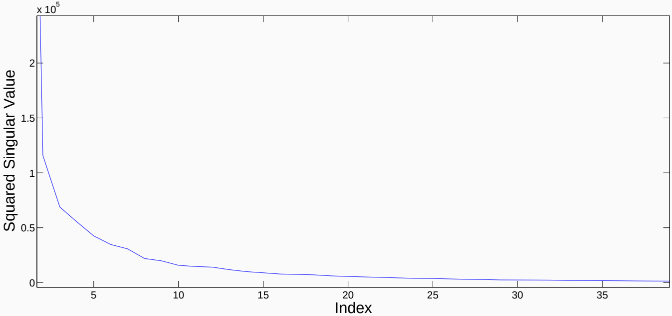
$$\|\mathbf{X}\mathbf{X}^\top \mathbf{A}_{\setminus k/\epsilon}\|_F^2 \le \epsilon \|\mathbf{A} - \mathbf{A}_k\|_F^2$$



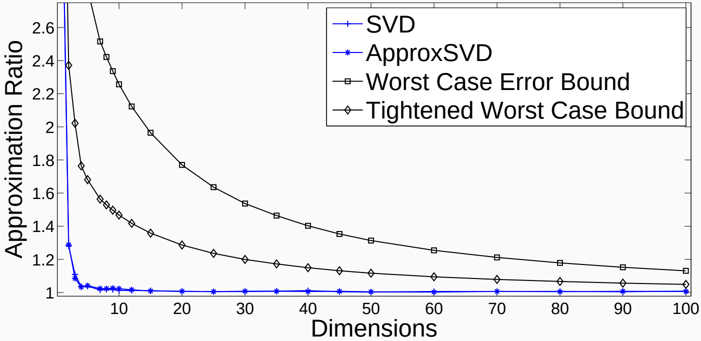
Recall that $\|\mathbf{B}\|_F^2 = \sum_i \sigma_i^2(\mathbf{B})$

- Analysis is very worst case since we assume $\sigma_{k/\epsilon+k}$ is just as big as σ_{k+1}.

- Analysis is very worst case since we assume $\sigma_{k/\epsilon+k}$ is just as big as σ_{k+1}.
- In practice, spectrum decay allows for a much smaller sketching dimension (Kappmeier, Schmidt, Schmidt '15).



- Analysis is very worst case since we assume $\sigma_{k/\epsilon+k}$ is just as big as σ_{k+1}.
- In practice, spectrum decay allows for a much smaller sketching dimension (Kappmeier, Schmidt, Schmidt '15).



All proofs have similar flavor:

$$\|\mathbf{A}_k - \mathbf{X}\mathbf{X}^\top \mathbf{A}_k\|_F^2 + \|\mathbf{A}_{\setminus k} - \mathbf{X}\mathbf{X}^\top \mathbf{A}_{\setminus k}\|_F^2$$

vs.

$$\|\mathbf{A}_k - \mathbf{X}\mathbf{X}^\top \mathbf{A}_k \boldsymbol{\Pi}\|_F^2 + \|\mathbf{A}_{\setminus k} - \mathbf{X}\mathbf{X}^\top \mathbf{A}_{\setminus k} \boldsymbol{\Pi}\|_F^2$$

All proofs have similar flavor:

$$\|A_k - XX^\top A_k\|_F^2 + \|A_{\setminus k} - XX^\top A_{\setminus k}\|_F^2$$

vs.

$$\|A_k - XX^\top A_k \Pi\|_F^2 + \|A_{\setminus k} - XX^\top A_{\setminus k}\Pi\|_F^2$$

Except that:

- We have to worry about cross terms $\|A\Pi - XX^\top A\Pi\|_F^2 \neq \|A_k\Pi - XX^\top A_k\Pi\|_F^2 + \|A_{\setminus k}\Pi - XX^\top A_{\setminus k}\Pi\|_F^2$

All proofs have similar flavor:

$$\|\mathbf{A}_k - \mathbf{X}\mathbf{X}^\top \mathbf{A}_k\|_F^2 + \|\mathbf{A}_{\setminus k} - \mathbf{X}\mathbf{X}^\top \mathbf{A}_{\setminus k}\|_F^2$$

vs.

$$\|\mathbf{A}_k - \mathbf{X}\mathbf{X}^\top \mathbf{A}_k \boldsymbol{\Pi}\|_F^2 + \|\mathbf{A}_{\setminus k} - \mathbf{X}\mathbf{X}^\top \mathbf{A}_{\setminus k} \boldsymbol{\Pi}\|_F^2$$

Except that:

- We have to worry about cross terms $\|\mathbf{A}\boldsymbol{\Pi} - \mathbf{X}\mathbf{X}^\top \mathbf{A}\boldsymbol{\Pi}\|_F^2 \neq \|\mathbf{A}_k\boldsymbol{\Pi} - \mathbf{X}\mathbf{X}^\top \mathbf{A}_k\boldsymbol{\Pi}\|_F^2 + \|\mathbf{A}_{\setminus k}\boldsymbol{\Pi} - \mathbf{X}\mathbf{X}^\top \mathbf{A}_{\setminus k}\boldsymbol{\Pi}\|_F^2$
- We have some hope of approximating $\|\mathbf{A}_{\setminus k} - \mathbf{X}\mathbf{X}^\top \mathbf{A}_{\setminus k}\|_F^2$

Rely on standard sketching tools:

- Approximate Matrix Multiplication: $\|\mathbf{A}\mathbf{\Pi}\mathbf{\Pi}^\top\mathbf{B}\|_F^2 \leq \epsilon\|\mathbf{A}\|_F\|\mathbf{B}\|_F$
- Subspace Embedding: $\|\mathbf{Y}\mathbf{A}\mathbf{\Pi}\|_F^2 = (1 \pm \epsilon)\|\mathbf{Y}\mathbf{A}\|_F^2$ *for all* $\mathbf{Y}$ if $\mathbf{A}$ has rank k
- Frobenius Norm Preservation: $\|\mathbf{A}\mathbf{\Pi}\|_F^2 = (1 \pm \epsilon)\|\mathbf{A}\|_F^2$

PROOF SKETCH FOR RANDOM PROJECTION

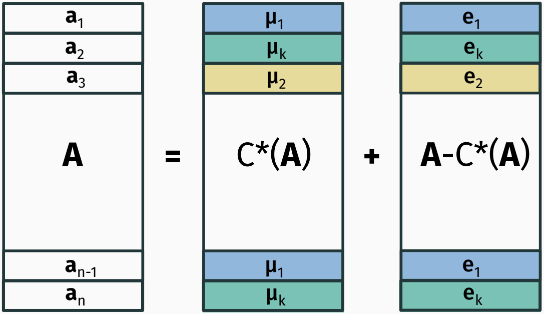
See paper for the details!

- Projection-cost preserving sketches guarantee $\|\tilde{\mathbf{A}} - \mathbf{X}\mathbf{X}^\top \tilde{\mathbf{A}}\|_F^2 \approx \|\mathbf{A} - \mathbf{X}\mathbf{X}^\top \mathbf{A}\|_F^2$ for all rank k $\mathbf{X}$.

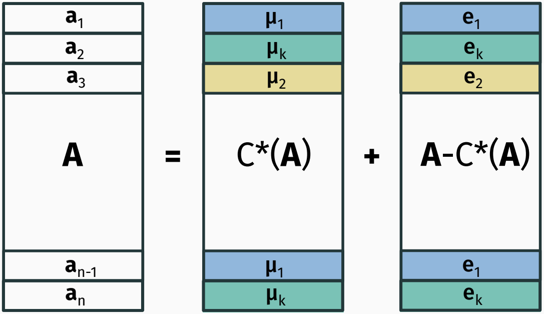
- Projection-cost preserving sketches guarantee $\|\tilde{\mathbf{A}} - \mathbf{X}\mathbf{X}^\top \tilde{\mathbf{A}}\|_F^2 \approx \|\mathbf{A} - \mathbf{X}\mathbf{X}^\top \mathbf{A}\|_F^2$ for all rank k $\mathbf{X}$.
- Standard proof approach: Split $\mathbf{A}$ into orthogonal pairs. Top of spectrum can be preserved multiplicatively, only top singular values of bottom of spectrum matter.

- Projection-cost preserving sketches guarantee $\|\tilde{A} - XX^\top \tilde{A}\|_F^2 \approx \|A - XX^\top A\|_F^2$ for all rank k X.
- Standard proof approach: Split A into orthogonal pairs. Top of spectrum can be preserved multiplicatively, only top singular values of bottom of spectrum matter.
- Dimensionality reduction for any constrained low rank approximation can be unified.

Split using *optimal* clustering:

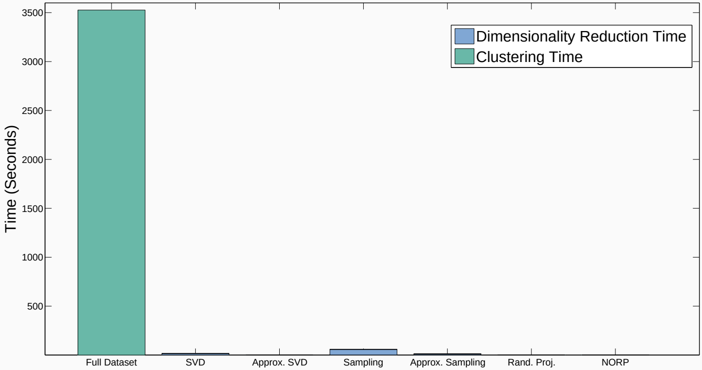


Split using *optimal* clustering:



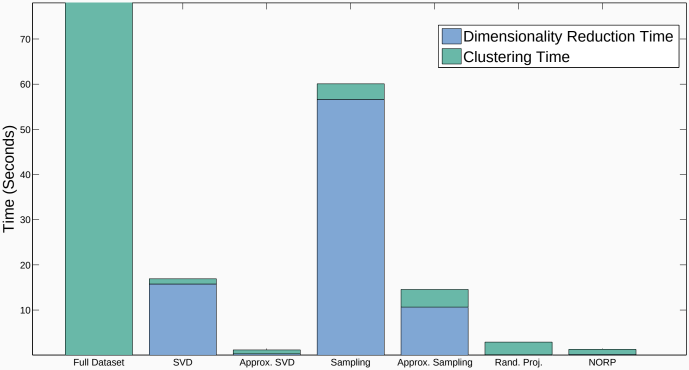
- $C^*(\mathbf{A})$ can be approximated with $O(\log k/\epsilon^2)$ dimensions.
- Not row orthogonal – have to use triangle inquality which leads to the $(9+\epsilon)$ factor.

Experiments & implements in [(Cameron) Musco '15]



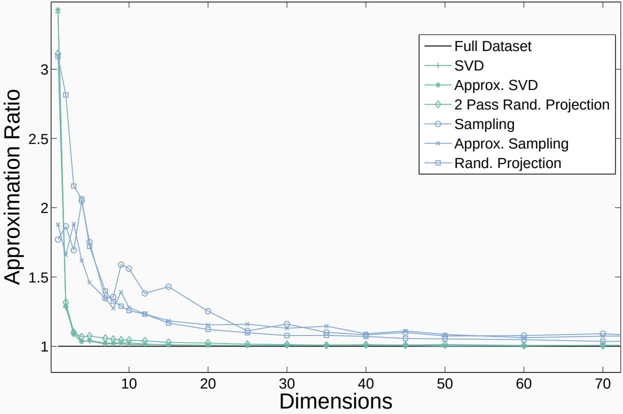
Time for $\epsilon = .01$ compared to baseline Lloyd's w/ k-means++.

Experiments & implements in [(Cameron) Musco '15]



Time for $\epsilon = .01$ compared to baseline Lloyd's w/ k-means++.

Experiments & implements in [(Cameron) Musco '15]



Thank you!

Open Questions:

- Improve our $O(\log k/\epsilon^2)$ random projection analysis from $(9+\epsilon)$ to $(1+\epsilon)$?
- Single pass PCA algorithm for turnstile streams with $1/\epsilon$ (instead of $1/\epsilon^2$) dependence?
- Coresets for k-means (reducing *number* of points instead of dimension) are difficult and messy. Can we get similarly "clean" analysis?