1) <u>Large Scale matrices</u>. $\leftarrow$ features

$$\begin{array}{c c} & 1 \quad 2 \quad \cdots j \cdots \quad n \\ \text{objects} \begin{array}{c} 1 \\ 2 \\ \vdots \\ i \\ \vdots \\ m \end{array} & \left[\begin{array}{ccc} & | & \\ \hline & \bullet & \\ & | & \end{array} \right]_{m \times n} \end{array}$$

A_{ij} = j^{th} feature of the i^{th} object

m - objects

n - features

. Geometrically: m <u>objects</u> in $\mathbb{R}^n$. $m < n$
points

1) <u>Astronomy</u>:

<u>objects</u>: Different angular regions in space

<u>Features</u>: Elements from frequency bands

2) <u>Genetics</u>

<u>objects</u>: List of genes

<u>Features</u>: Level A_{ij} of the i^{th} gene in j^{th} individu

3) <u>Document / Text</u>

<u>objects</u>: List of documents

<u>Features</u>: A_{ij} = frequency of j^{th} term in i^{th} document

4) Discretization of ODE / PDE

5) Kernel matrices when describing pairwise relations

<u>Properties of these matrices</u>

. Singular values decay quickly suggesting the potential to compress or reduce dimension

Finding structures with randomness

. Two step process

1) construct a low-dimensional subspace that captures the action of a matrix

2) Project / restrict the matrix to this subspace and process that matrix using standard factorization - QR, SVD, CUR of the resulting reduced matrix.

Stage 1: Compute an approximate basis for the range of the input matrix $A|_{m\times n}$ consisting of columns of Q which are orthonormal: $Q_{m\times r}$:

$$A \simeq Q Q^T A \longrightarrow (1)$$

where r is small but large enough to capture the actions of A

stage 2: Given Q that satisfies (1), use Q to get standard factorization of A.

. Stage 1 is implemented by randomized approach - random sampling and random projection

. Stage 2 uses std. deterministic approach

. This separation is key to the success of this approach.

An example: SVD computation

1) From (1): $A_{m\times n} \simeq Q Q^T A$. SET $B_{n\times n} = Q^T A \rightarrow (2)$

$Q_{m\times n}$ - n is the rank of Approx.

$Q^T Q = I_n$, $Q Q^T$ - Projection

This gives a low-rank factorization

$$A \simeq Q B$$

2) Compute SVD of $B = \tilde{U} \Sigma V^T$

3) Set $U = Q\tilde{U}$ and $A \simeq Q B$

$$= Q \tilde{U} \Sigma V^T$$
$$= U \Sigma V^T$$

Two ways

1) Fixed Precision Approximation

Statement

- Given A and $\varepsilon > 0$, seek a Q with ~~[illegible]~~ orthogonal columns and $k = k(\varepsilon)$:

$$\| A - Q Q^T A \| \leq \varepsilon \longrightarrow (3)$$

- Dimension k of the range of Q captures the action of A.

Pathway to solve

- SVD furnishes an optimal answer to this fixed precision problem. For each j

$$\min_{\text{rank}(X) \leq j} \| A - X \| \leq \sigma_{j+1} \longrightarrow (4)$$

- Choosing $X = Q Q^T A$ where $Q_{m\times k}$ has k orthogonal left singular vectors of A which guarantees (3)

2) **Fixed rank Approx**: Given A and k and an over sampling parameter $p > 0$, Construct Q with $(k+p)$ orthonormal columns

$$\|A - QQ^T A\| = \min_{\text{rank}(X) \le k} \|A - X\| \rightarrow (5)$$

Remark: How randomness helps to solve fixed rank problem?

- Seek a basis for Range(A) with rank k:
- Let w be a random vector. Then $y = Aw$ is a random sample from Range(A)
- Build $Y = \{y_1, y_2 \cdots y_k\}$: $y_i = A w_i$
- This set Y is in general position.
- Then orthogonalize Y.

Note: Let $A = B + E$ $\begin{cases} B - \text{rank } k \\ E - \text{Perturbation} \end{cases}$

oversampling $\boxed{w_i \in \mathbb{R}^n}$

$$y_i = A w_i = B w_i + E w_i \quad 1 \le i \le k+p \rightarrow (6)$$

- Goal is to find a basis that covers B
- E shifts the direction sample vectors outside of the range of B which can prevent the Span(Y) to cover the Range(B)
- The enriched set in (6) can reduce the leak outside of Range(B)

· Small p helps $p = 5$ to 10.

· This is the basis for randomized algorithm.

Fixed Rank Algorithm: Proto-type

Given: $A \in \mathbb{R}^{m \times n}$ Target Rank $= k$,
over sampling parameter p.

output: $Q\big|_{m \times (k+p)}$

1) Draw $n \times (k+p)$ random matrix $\Omega\big|_{n \times (k+p)}$

2) Compute $\underset{m \times (k+p)}{Y} = \underset{m \times n}{A}\, \underset{n \times (k+p)}{\Omega} = [Y_1\ Y_2 \cdots Y_{k+p}]$

3) orthonormalize the columns of $Y = QR$ and get $Q\big|_{m \times (k+p)}$. This is the basis for Range (A)

. Gram-schmidt
. House holder
. Givens

Performance Analysis:

Theorem 1.1 $A \in \mathbb{R}^{m \times n}$, Select a target rank $k \geq 2$ and an oversize parameter $p \geq 2$ where $(k+p) < \min\{m, n\}$. Execute the prototype algorithm with $\Omega = [\Omega_{ij}]$, $\Omega_{ij} \sim$ iid $N(0,1)$ to obtain $Q\big|_{m \times (k+p)}$ with orthonormal columns.

Then

$$\underset{\substack{\downarrow \\ (\text{w.r.to random } \Omega)}}{E} \| A - QQ^T A \| \leq \left\{ \left[1 + \frac{4\sqrt{k+p}}{p-1} \right] (m \wedge n)^{1/2} \right\} \sigma_{k+1} \quad \rightarrow (7)$$

. known as a sharper than deterministic ~~analog~~ analog based on RRQR Algorithm of Gu and Eisenstat (1996)

. Thanks to measure concentration: the following holds:

$$\| A - QQ^T A \| \leq \left[1 + 11 \sqrt{k+p} \cdot (m \wedge n)^{1/2} \right] \sigma_{k+1}$$

~~holds~~ with probability at least $(1 - 6p^{-p})$. $p = 5$ gives good result.

Historical facts.

Existence

1) Column selection (Existence): ~~For~~ Every $A \in \mathbb{R}^{m \times n}$ Contains a ~~submatrix~~ C with k ~~columns~~ Columns s.t.

$$\| A - CC^+ A \| \leq \sqrt{1 + k(n-k)} \, \| A - A_{(k)} \| \quad \rightarrow (8)$$

where $A_{(k)}$ is the ~~best~~ k-rank approximation to A.

A. F. Ruston (1964) "Auerbach's Theorem", Math. Proc. Cambridge Phil. Society vol 56 (1964) pp 476-480

Note: . Column selection is NP-hard.

. Efficient deterministic RRQR method will achieve the above bound in (8)

Column selection

2) Using this there exists randomized algorithm for fixed rank approximation:

$$\| A - QQ^T A \| \leq \min_{\text{rank}(X) \leq k} \| A - X \| \quad \rightarrow (9)$$

- This approach use sampling probabilities based on
  - (1) Squared 2-norms of rows / columns
  - (2) Leverage score that reflects the relative importance of columns
- Columns are selected using this distribution

- Earliest method for randomized column selection is by Frieze, Kannan, Vempala (1998)

FKV (1998)

- Given A and $l(\varepsilon, k)$ select B:

$$\|A - B\|_F \le \|A - A_{(k)}\|_F + \varepsilon \|A\|_F$$

Rudelson + Vershynin (2007)

- using the same sampling, proved $\|\cdot\|_2$ bounds

- **Deshpande et al (2006)**

$$\|A - B\|_F \le (1+\varepsilon)\|A - A_{(k)}\|_F$$

3 Dimension reduction using random Projection

- Started in 1984 - Johnson-Lindenstrauss
- l_2 embedding

Summary:-

- Stage 1 is realized in one of 3-ways

(1)
$y_i = A w_i$
$Y = [y_1 \dots y_k]$
$Y = A\Omega$

(2)
Importance Sampling

(3)
Random Projection
Somewhat similar to (1).

Sampling based randomized algorithms

Fact $x \in \mathbb{R}^n$ a random vector

- $E(x) \in \mathbb{R}^n \longrightarrow (1)$
- $\text{Var}(x) = E\left[\|x - E(x)\|^2\right]$

$$= E\left[x - E(x)\right]^T \left[x - E(x)\right]$$

$$= \sum_i E\left(x_i - E(x_i)\right)^2 \text{ - Total Variance}$$

- $\text{cov}(x) = E\left[x - E(x)\right]\left[x - E(x)\right]^T \longrightarrow (2)$

$$E\left[xx^T - x^T E(x) - E^T(x)\, x + E^T(x) E(x)\right]$$

$$= E\left[x x^T\right] - E^T(x) E(x)$$

$$= E\left[\|x\|^2\right] - \|E(x)\|^2 \quad (2a)$$

- $\text{Var}(x) = \text{tr}\left[\text{cov}(x)\right] \longrightarrow (3)$

$\boxed{\text{Problem 1}}$ **Matrix-Vector product**

- $A \in \mathbb{R}^{m \times n}$, $v \in \mathbb{R}^n$, $Av = \sum_j A_{*j} v_j \longrightarrow (4)$

= sum of n vectors.

- (Av) can be estimated using a sample of column vectors of A
- Error in the estimate is measured by the variance of the estimate.
- Two ways to sample the columns:
  - UNIFORM DISTRIBUTION
  - Importance distribution

Motivating Exercise 1 **Scaled random variable**

A case for data dependent sampling

. Let x be a r.v. and $\boxed{x = \frac{a_i}{p_i}}$ with prob $= p_i$

with $\sum_{i=1}^{n} p_i = 1$. Let $n=2$ for simplicity

$$\cdot\ E(x) = \frac{a_1}{p_1}\cdot p_1 + \frac{a_2}{p_2}\cdot p_2 = a_1 + a_2 \triangleq \bar{a}$$

$$\cdot\ \mathrm{Var}(x) = p_1\left[\frac{a_1}{p_1} - \bar{a}\right]^2 + p_2\left[\frac{a_2}{p_2} - \bar{a}\right]^2$$

$$= p_1\left[\frac{a_1^2}{p_1^2} - \frac{2a_1\bar{a}}{p_1} + \bar{a}^2\right] + p_2\left[\frac{a_2^2}{p_2^2} + \bar{a}^2 - \frac{2a_2\bar{a}}{p_2}\right]$$

$$= \frac{a_1^2}{p_1} + \frac{a_2^2}{p_2} + (p_1+p_2)\bar{a}^2 - (2a_1\bar{a} + 2a\bar{a})$$

$$= \frac{a_1^2}{p_1} + \frac{a_2^2}{p_2} + \underbrace{\bar{a}^2 - 2\bar{a}(a_1+a_2)}$$

$$\left[\begin{aligned}&(a_1+a_2)^2\\ &-2(a_1+a_2)(a_1+a_2)\\ &= -(a_1+a_2)^2\end{aligned}\right.$$

$$= \frac{a_1^2}{p_1} + \frac{a_2^2}{p_2} - (a_1+a_2)^2$$

Special case Set $p_1 = p_2 = 1/2$ – uniform sampling

$$E(x) = a_1 + a_2$$

$$\mathrm{Var}(x) = 2a_1^2 + 2a_2^2 - (a_1^2 + 2a_1a_2 + a_2^2)$$

$$= a_1^2 - 2a_1a_2 + a_2^2 = (a_1-a_2)^2 \geq 0$$

. This is higher than the sampling below.

---

optimal p_1 & p_2 What choice of p_i's will min. Var(x)?

$$\mathrm{Min}\left(\frac{a_1^2}{p_1} + \frac{a_2^2}{p_2}\right) \text{ when } p_1 + p_2 = 1$$

$$L(p,\lambda) = \frac{a_1^2}{p_1} + \frac{a_2^2}{p_2} + \lambda(p_1 + p_2 - 1)$$

$$\frac{\partial L}{\partial p_1} = -\frac{a_1^2}{p_1^2} + \lambda = 0 \Rightarrow \lambda = \frac{a_1^2}{p_1^2} \Rightarrow \boxed{p_1^2 = \frac{a_1^2}{\lambda}} \ldots$$

$$\frac{\partial L}{\partial p_2} = -\frac{a_2^2}{p_2^2} + \lambda = 0 \Rightarrow \lambda = \frac{a_2^2}{p_2^2} \Rightarrow \boxed{p_2^2 = \frac{a_2^2}{\lambda}}$$ (3)

- But $p_1 + p_2 = 1 \Rightarrow \frac{a_1}{\sqrt{\lambda}} + \frac{a_2}{\sqrt{\lambda}} = 1 \Rightarrow a_1 + a_2 = \sqrt{\lambda}$

$\therefore$ optimal $p_1^* = \frac{a_1}{a_1+a_2}$ $\quad p_2^* = \frac{a_2}{a_1+a_2}$

$$\text{Var}(x) = \frac{a_1^2}{p_1} + \frac{a_2^2}{p_2} - (a_1+a_2)^2$$

$$= \frac{a_1^2(a_1+a_2)}{a_1} + \frac{a_2^2(a_1+a_2)}{a_2} - (a_1+a_2)^2$$

$$= a_1(a_1+a_2) + a_2(a_1+a_2) - (a_1+a_2)^2$$

$$= (a_1+a_2)^2 - (a_1+a_2)^2 = 0$$

$\therefore$ variance reduction is possible using a data dependent sampling distribution

---

- <u>Return to Matrix-Vector multiply</u>

- What is a good sampling distribution for the columns of A?

key step

- Define a random vector $x = \frac{A_{*j} v_j}{p_j}$ with probability p_j, $1 \le j \le n$

- <u>scaled random vector scaled by p_j</u>

- $E(x) = \sum_j p_j \frac{A_{*j} v_j}{p_j} = \sum A_{*j} v_j = Av \quad \rightarrow (4)$

(ii) x is an <u>unbiased random vector</u>

$$\text{Var}(X) = E\|X\|^2 - \|E(X)\|^2$$ from (2a) in page (1)

$$= \sum_j \frac{\|A_{*j} V_j\|^2}{p_j^2} p_j - \|AV\|^2$$ From (4)

$(\because \|a x\| = |a| \, \|x\|)$

Similar to the example

$$= \sum_j \frac{N_j^2 \|A_{*j}\|^2}{p_j} - \underbrace{\|AV\|^2}_{\text{Constant w.r.to } p_j} \longrightarrow (5)$$

- choice of p_j depends on the first term on the r.h.s. of (5).
- $LS_{col}(A)$ - [Distribution that depends on the squared length of columns of A

choice of p_j

- column j is picked with prob. $\} = p_j = \dfrac{\|A_{*j}\|^2}{\sum_j \|A_{*j}\|^2}$

$$= \frac{\|A_{*j}\|^2}{\|A\|_F^2} \longrightarrow (6)$$

- Substitute (6) in (5)

$$\text{Var}(X) = \sum_j N_j^2 \frac{\|A_{*j}\|^2}{\|A_{*j}\|^2} \cdot \|A\|_F^2 - \|AV\|^2$$

$$= \|A\|_F^2 \sum_j N_j^2 - \|AV\|^2$$

$$= \|A\|_F^2 \|N\|_2^2 - \|AV\|^2$$

$$\le \|A\|_F^2 \|N\|_2^2 \longrightarrow (7)$$

useful when $\|AV\|$ is comparable to $\|A\|_F \|N\|_2$

$$LS_{sol}(A) = \{ p_1, p_2, \cdots p_n \mid \sum p_j = 1, \ p_j \text{ is in } (6) \}$$

- To approximate AV:
  - Perform s trials (with replacement) and take the average

$$Y = \frac{1}{s} \sum_{t=1}^{s} \left[\frac{A_{*j_t} \, V_{j_t}}{p_{j_t}} \right] \approx AV$$

$$= \frac{1}{s} \sum_{t=1}^{s} x_t = \sum_t \left(\frac{x_t}{s} \right)$$

$$\therefore \; E(Y) = \frac{1}{s} \sum_{t=1}^{s} E(x_t) = \frac{s(AV)}{(s)} = AV$$

- $$\mathrm{Var}(Y) = \frac{1}{s^2} \sum_t \mathrm{Var}(x_t) \qquad [\text{using } (7)]$$

$$\leq \frac{1}{s^2} \cdot \left[\sum_{j=1}^{m} \|A\|_F \|V\|_2^2 \right]$$

$$= \frac{1}{s} \|A\|_F^2 \|V\|_2^2 \longrightarrow (8)$$

Note:-

(1) We could row sampling as well using a similar argument to get $RS(A)$ distribution.

<u>Problem 2</u> <u>MATRIX - MATRIX Product</u>

$$C = A\Big|_{m\times n} \; B\Big|_{n \times k}$$ — $\boxed{\text{Three ways of multiplying Matrices}}$

$$C = AB = (AB_{*1}, AB_{*2}, \dots AB_{*k})$$

- Apply Matrix-Vector product routine k times and collate the result.

<u>claim</u>: [n is the common # of columns of A and rows of B.

- Define p_i : $\sum_{i=1}^{n} p_i = 1$

- Define a random matrix $\underbrace{X_j}_{\text{with } X =} = \dfrac{A_{*j} B_{j*}}{p_j}$ → (1)

  with probability p_j, $1 \le j \le n$

- $$E(X) = \sum_{j=1}^{n} X_j p_j = \sum_{j=1}^{n} A_{*j} B_{j*} = AB$$ → (2)

- $$p_j = \frac{\|A_{*j}\|^2 \|B_{j*}\|^2}{\sum_{j=1}^{n} \|A_{*j}\|^2 \|B_{j*}\|^2} \quad 1 \le j \le n$$ → (3)

- Define $$Y = \frac{1}{s} \sum_{t=1}^{s} X_t = \frac{1}{s} \sum_{t=1}^{s} \left(\frac{A_{*j_t} B_{j_t *}}{p_{j_t}} \right)$$

- $E(Y) = AB$ => unbiased → (4)

- $$Var(Y) = \frac{1}{s} \sum_{t=1}^{s} \frac{\|A_{*j_t}\|^2 \|B_{j_t *}\|^2}{p_j} - \|AB\|_F^2$$ → (5)

  $\boxed{\text{using } p_j \text{ as in (3)}}$

- $$\le \frac{1}{s} \|A\|_F^2 \|B\|_F^2$$ → (6)

<u>Implementation</u>

- Let $k_1, k_2, \ldots k_s$ be s integers chosen in s trials.
- Then,

$$\frac{1}{s}\sum x_i = \frac{1}{s}\left[\frac{A_{*k_1}B_{k_1*}}{p_{k_1}} + \frac{A_{*k_2}B_{k_2*}}{p_{k_2}} + \cdots \frac{A_{*k_s}B_{k_s*}}{p_{k_s}}\right]$$

$$= [A_{*k_1}, A_{*k_2}, \cdots A_{*k_s}] \begin{bmatrix} B_{k_1*}/sp_{k_1} \\ B_{k_2*}/sp_{k_2} \\ \vdots \\ B_{k_s*}/sp_{k_s} \end{bmatrix}$$

$$= \underbrace{C}_{\substack{\text{chosen} \\ \text{columns} \\ \text{of } A}} \; \tilde{B}$$

$\tilde{B}$ = chosen rows of B scaled by sp_k

$$\underset{m \times n}{[A]}\;\underset{n \times p}{[B]} \simeq \underset{m \times s}{[C]}\;\underset{s \times p}{[\tilde{B}]} \quad \longrightarrow (7)$$

- Error <u>bound is given by</u> (6) above.

<u>Special cases:</u>

1) If $B = A^T$, we can approximate the Gramian AA^T and hence the singular values of A.

2) When $B = A^T$: start with

$$\mathrm{Var}(Y) \le \frac{1}{s} \sum \frac{\|A\|_F^2 \, \|A^T\|_F^2}{p_j} - \|AA^T\|_F^2$$

$$= \underbrace{\frac{1}{s} \sum \frac{\|A\|_F^4}{p_j}}_{\text{minimize w.r.to } p_j} - \|AA^T\|_F^2 \quad \to (8)$$

<u>Exercise 2</u> Let $a_1, a_2, \ldots a_n$ be + real numbers. Prove that $\sum_{j=1}^{n} \frac{a_j}{x_j}$ when $\sum x_j = 1$ attains the minimum when $x_j = \frac{\sqrt{a_j}}{\sum \sqrt{a_j}}$. (See Exercise 1)

sketch

Problem 3. Low rank approximation for A is called sketch of A, where the columns and rows each picked in ~~[illegible]~~ individual trials based on importance distribution based on squared lengths of columns / rows is a good sketch. Let $A \in \mathbb{R}^{m \times n}$

1) Pick s columns of A and form $C|_{m \times s}$ based on $LS_{col}(A)$
2) Pick r rows of A from $LS_{row}(A)$ and form $R_{s \times n}$
3) Using C and R express: $A \approx CUR$ by computing U

A comparison with SVD

Note:- . You may think of using SVD to get a good sketch. While it is true, it takes a long time: $A = U \Delta V^T$

. Besides, in SVD the ~~x~~columns of U and rows of V are ~~[illegible] combinations of those of A~~ not directly ~~related to those of A~~ from A but are linear combinations.

. In here, we directly pick the columns and row of A by sampling and is called interpolative approximation

. SVD gives best approximation, the CUR is not optimal in that sense.

How to choose U? $A \in \mathbb{R}^{m \times n}$

1) An Intuition: $A = AI$

- Sample s columns of A and get $C|_{m \times s}$
- Let W be the corresponding s rows of I scaled as in Matrix-Matrix product.
- $A = AI = CW \longrightarrow (1)$
- From (b): $E\|AI - CW\|^2 \le \dfrac{\|A\|_F^2 \|I\|_F^2}{s}$

$$= \|A\|_F^2 \left(\frac{n}{s}\right) \rightarrow (2)$$

- Since we need $s > n$, this is not viable, but the idea stands.

2) Modification: $A \in \mathbb{R}^{m \times n}$

- Consider $C|_{m \times s}$ and $R_{r \times n}$ that we selected earlier, where $s < m$, $r < n$.
- If R is of full-rank, then

$RR^T|_{r \times r}$ is SPD

$R^+|_{n \times r} = R^T (RR^T)^{-1}$ is the Moore-Penrose inverse

$P|_{n \times n} = R^T (RR^T)^{-1} R$ is the projection onto the row space of R.

- Recall P acts as identity on the ~~row sp~~ Row space of R.

Illustration eg: $R = (1, 1)$ $RR^T = (1,1)\begin{pmatrix}1\\1\end{pmatrix} = 2$

$$R^+ = R^T (RR^T)^{-1} = \frac{1}{2}\begin{pmatrix}1\\1\end{pmatrix}$$

$$P = R^T (RR^T)^{-1} R = \frac{1}{2}\begin{pmatrix}1\\1\end{pmatrix}(1,1) = \frac{1}{2}\begin{pmatrix}1 & 1\\1 & 1\end{pmatrix}$$

$\boxed{\text{Row space of } R}$ $y = R^T a \Rightarrow Py = R^T (RR^T)^{-1} R R^T a$

$a = \begin{pmatrix}a_1\\a_2\end{pmatrix}$ $= R^T a = y$

- So we can replace I by P in the above analysis.

- $AP = A \underbrace{R^T(RR^T)^{-1}} R \approx CUR$ $\boxed{U = R^T(RR^T)^{-1}}$

$\therefore$ By the inequality (6):

$\boxed{\|A\|_2^2 \le \|A\|_F^2}$

$$E\left[\|AP - CUR\|_2^2\right] \le E\left(\|AP - CUR\|_F^2\right)$$

$$\le \frac{\|A\|_F^2 \|P\|_F^2}{s} = \|A\|_F^2\left(\frac{r}{s}\right) \rightarrow (3)$$

where $\|P\|_F^2 = r$. (why?)

$\boxed{\begin{array}{l} R_{r\times m} \\ R^+_{n\times r} \\ P_{m\times m} \\ (RR^T)_{r\times r} \end{array}}$

- Recall $\|A\|_F^2 = tr(AA^T)$

$$\Rightarrow \|P\|_F^2 = tr[P] = tr[R^T(RR^T)^{-1}R]$$

$\boxed{tr(AB) = tr(BA)}$

$$= tr[RR^T(RR^T)^{-1}]$$

$$= tr[I_r] = r \rightarrow (4)$$

Combining:

$$A - CUR = A - AP + AP - CUR$$

Triangle inequality

$$\|A - CUR\|_2 \le \|A - AP\|_2 + \|AP - CUR\|_2$$

Recall: $x \le y + z$

$$\Rightarrow x^2 \le (y + z)^2 = \underbrace{y^2 + 2yz + z^2 < 2(y^2 + z^2)}$$

$$\Rightarrow 0 \le (y^2 - 2yz + z^2) = (y-z)^2$$

$$\therefore \|A - CUR\|_2^2 \le 2\|A - AP\|_2^2 + 2\|AP - CUR\|_2^2$$