Randomized low rank approximation

L. Grigori

Inria Paris, UPMC

November 2021
Plan

Randomization for least-squares problem

Low rank matrix approximation

Randomized algorithms for low rank approximation
Plan

Randomization for least-squares problem

Low rank matrix approximation

Randomized algorithms for low rank approximation
Johnson-Lindenstrauss transform

Definition 3 from [Woodruff, 2014].
A random matrix $\Omega_1 \in \mathbb{R}^{k \times m}$ is a Johnson-Lindenstrauss transform with parameters $\epsilon, \delta, n$, or $\text{JLT}(n, \epsilon, \delta)$, if with probability at least $1 - \delta$ for any $n$-element subset $V \subset \mathbb{R}^m$, for all $x_i, x_j \in V$, we have

$$\left| \langle \Omega_1 x_i, \Omega_1 x_j \rangle - \langle x_i, x_j \rangle \right| \leq \epsilon \|x_i\|_2 \|x_j\|_2$$

(1)

- If $x_i = x_j$ we obtain $\|\Omega_1 x_i\|_2^2 = (1 \pm \epsilon)\|x_i\|_2^2$.
- It can also be expressed as: given all vectors $x_i, x_j \in V$ are rescaled to be unit vectors, then for all $x_i, x_j \in V$ we require to hold:

$$\|\Omega_1 x_i\|_2^2 = (1 \pm \epsilon)\|x_i\|_2^2$$

(2)

$$\|\Omega_1 (x_i + x_j)\|_2^2 = (1 \pm \epsilon)\|x_i + x_j\|_2^2$$

(3)

Proof that we obtain relation (4):

$$\langle \Omega_1 x_i, \Omega_1 x_j \rangle = \left( \|\Omega_1 (x_i + x_j)\|_2^2 - \|\Omega_1 x_i\|_2^2 - \|\Omega_1 x_j\|_2^2 \right) / 2$$

$$= \left( (1 \pm \epsilon)\|x_i + x_j\|_2^2 - (1 \pm \epsilon)\|x_i\|_2^2 - (1 \pm \epsilon)\|x_j\|_2^2 \right) / 2$$

$$= \langle x_i, x_j \rangle \pm O(\epsilon)$$
Let $\Omega_1 \in \mathbb{R}^{k \times m}$ be a matrix whose entries are independent standard normal random variables, multiplied by $1/\sqrt{k}$. If $k = O(\epsilon^{-2} \log (n/\delta))$, then $\Omega_1$ is a JLT$(n, \epsilon, \delta)$.

Source: Theorem 4 in [Woodruff, 2014], see also Theorem 2.1 and proof in S. Dasgupta, A. Gupta, 2003, An Elementary Proof of a Theorem of Johnson and Lindenstrauss
Let $\Omega_1 \in \mathbb{R}^{k \times m}$ be a matrix whose entries are independent standard normal random variables, multiplied by $1/\sqrt{k}$. If $k = O(\epsilon^{-2}(n + \log (1/\delta)))$, then $\Omega_1$ is an oblivious subspace embedding (OSE) with parameters $(n, \epsilon, \delta)$. That is, with probability at least $1 - \delta$ for any $n$-dimensional subspace $V \subset \mathbb{R}^m$, for all $x_i, x_j \in V$, we have

$$|\langle \Omega_1 x_i, \Omega_1 x_j \rangle - \langle x_i, x_j \rangle| \leq \epsilon \|x_i\|_2 \|x_j\|_2 \quad (4)$$

Source: Theorem 6 in [Woodruff, 2014]
Least squares problems

Given $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^n$, with $m \ll n$, solve

$$y := \arg \min_{x \in \mathbb{R}^n} \|Ax - b\|_2$$

1. Solve by computing QR factorization of $A$ or using normal equations,

$$A^T A x = A^T b.$$

2. Solve by using randomization, with $\Omega_1 \in \mathbb{R}^{k \times m}$

$$y^* := \arg \min_{x \in \mathbb{R}^n} \|\Omega_1 (Ax - b)\|_2$$
Solve by using randomization, with $\Omega_1 \in \mathbb{R}^{k \times m}$, $k = O(\epsilon^{-2}(n + \log(1/\delta)))$, being OSE with parameters $(n, \epsilon, \delta)$ for $V = \text{range}(A) + \text{span}(b)$

$$y^* := \arg \min_{x \in \mathbb{R}^n} \|\Omega_1(Ax - b)\|_2$$

We obtain with probability $1 - \delta$:

$$\|Ay^* - b\|_2^2 \leq (1 + O(\epsilon))\|Ay - b\|_2^2$$
Plan

Randomization for least-squares problem

Low rank matrix approximation

Randomized algorithms for low rank approximation
Low rank matrix approximation

- Problem: given $A \in \mathbb{R}^{m \times n}$, compute rank-$k$ approximation $ZW^T$, where $Z$ is $m \times k$ and $W^T$ is $k \times n$.

- Problem with diverse applications
  - from scientific computing: fast solvers for integral equations, H-matrices
  - to data analytics: principal component analysis, image processing, ...

\[
Ax \rightarrow ZW^T x
\]

\[
\text{Flops} \quad 2mn \rightarrow 2(m + n)k
\]
Singular value decomposition

For any given \( A \in \mathbb{R}^{m \times n} \), \( m \geq n \) its singular value decomposition is

\[
A = U\Sigma V^T = (U_1 \quad U_2 \quad U_3) \cdot \begin{pmatrix}
\Sigma_1 & 0 \\
0 & \Sigma_2 \\
0 & 0
\end{pmatrix} \cdot (V_1 \quad V_2)^T
\]

where

- \( U \in \mathbb{R}^{m \times m} \) is orthogonal matrix, the left singular vectors of \( A \),
  - \( U_1 \) is \( m \times k \), \( U_2 \) is \( m \times n - k \), \( U_3 \) is \( m \times m - n \)
- \( \Sigma \in \mathbb{R}^{m \times n} \), its diagonal is formed by \( \sigma_1(A) \geq \ldots \geq \sigma_n(A) \geq 0 \)
  - \( \Sigma_1 \) is \( k \times k \), \( \Sigma_2 \) is \( n - k \times n - k \)
- \( V \in \mathbb{R}^{n \times n} \) is orthogonal matrix, the right singular vectors of \( A \),
  - \( V_1 \) is \( n \times k \), \( V_2 \) is \( n \times n - k \)
Properties of SVD

Given $A = U\Sigma V^T$, we have

- $A^T A = V \Sigma^T \Sigma V^T$, the right singular vectors of $A$ are a set of orthonormal eigenvectors of $A^T A$.
- $AA^T = U \Sigma^T \Sigma U^T$, the left singular vectors of $A$ are a set of orthonormal eigenvectors of $AA^T$.
- The non-negative singular values of $A$ are the square roots of the non-negative eigenvalues of $A^T A$ and $AA^T$.
- If $\sigma_k \neq 0$ and $\sigma_{k+1}, \ldots, \sigma_n = 0$, then
  $\text{Range}(A) = \text{span}(U_1)$, $\text{Null}(A) = \text{span}(V_2)$,
  $\text{Range}(A^T) = \text{span}(V_1)$, $\text{Null}(A) = \text{span}(U_2 U_3)$. 
Norms

\[ \|A\|_p = \max_{\|x\|_p = 1} \|Ax\|_p \]

\[ \|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2} = \sqrt{\sigma_1^2(A) + \ldots + \sigma_n^2(A)} \]

\[ \|A\|_2 = \sigma_{\text{max}}(A) = \sigma_1(A) \]

Some properties:

\[ \max_{i,j} |A(i,j)| \leq \|A\|_2 \leq \sqrt{mn} \max_{i,j} |A(i,j)| \]

\[ \|A\|_2 \leq \|A\|_F \leq \sqrt{\min(m,n)} \|A\|_2 \]

Orthogonal Invariance: If \( Q \in \mathbb{R}^{m \times m} \) and \( Z \in \mathbb{R}^{n \times n} \) are orthogonal, then

\[ \|QAZ\|_F = \|A\|_F \]
\[ \|QAZ\|_2 = \|A\|_2 \]
Low rank matrix approximation

Best rank-$k$ approximation $A_{opt,k} = U_k \Sigma_k V_k$ is rank-$k$ truncated SVD of $A$ [Eckart and Young, 1936]

$$\min_{\text{rank}(A_k) \leq k} \|A - A_k\|_2 = \|A - A_{opt,k}\|_2 = \sigma_{k+1}(A)$$  \hspace{1cm} (5)

$$\min_{\text{rank}(A_k) \leq k} \|A - A_k\|_F = \|A - A_{opt,k}\|_F = \sqrt{\sum_{j=k+1}^{n} \sigma_j^2(A)}$$  \hspace{1cm} (6)

Image source: https://pixabay.com/photos/billiards-ball-play-number-half-4345870/
Matrix $A$ might not exist entirely at a given time, rows or columns are added progressively.

- Streaming algorithm: can solve an arbitrarily large problem with one pass over the data (a row or a column at a time).
- Weakly streaming algorithm: can solve a problem with $O(1)$ passes over the data.

Matrix $A$ might exist only implicitly, and it is never formed explicitly.
Low rank matrix approximation: trade-offs

Communication optimal if computing a rank-k approximation on $P$ processors requires

$$\# \text{ messages} = \Omega \left( \log_2 P \right).$$
Communication optimal if computing a rank-k approximation on $P$ processors requires

$$\# \text{ messages} = \Omega (\log_2 P).$$
Idea underlying many algorithms

Compute $\tilde{A}_k = PA$, where $P = P^o$ or $P = P^{so}$ is obtained as:

1. Construct a low dimensional subspace $X = \text{range}(A\Omega_1)$, $\Omega_1 \in \mathbb{R}^{n \times l}$ that approximates well the range of $A$, e.g.

   $\|A - P^o A\|_2 \leq \gamma \sigma_{k+1}(A)$, for some $\gamma \geq 1$,

   where $Q_1$ is orth. basis of $(A\Omega_1)$

   $$P^o = A\Omega_1 (A\Omega_1)^+ = Q_1 Q_1^T$$, or equiv $P^o a_j := \arg \min_{x \in X} \|x - a_j\|_2$

2. Select a semi-inner product $\langle \Theta_1 \cdot, \Theta_1 \cdot \rangle_2$, $\Theta_1 \in \mathbb{R}^{l' \times m}$ $l' \geq l$, define

   $P^{so} = A\Omega_1 (\Theta_1 A\Omega_1)^+ \Theta_1$, or equiv $P^{so} a_j := \arg \min_{x \in X} \|\Theta_1 (x - a_j)\|_2$
Idea underlying many algorithms

Compute $\tilde{A}_k = PA$, where $P = P^o$ or $P = P^{so}$ is obtained as:

1. Construct a low dimensional subspace $X = \text{range}(A\Omega_1)$, $\Omega_1 \in \mathbb{R}^{n \times l}$ that approximates well the range of $A$, e.g.

$$\|A - P^o A\|_2 \leq \gamma \sigma_{k+1}(A), \text{ for some } \gamma \geq 1,$$

where $Q_1$ is orth. basis of $(A\Omega_1)$

$$P^o = A\Omega_1 (A\Omega_1)^+ = Q_1 Q_1^T,$$

or equiv $P^o a_j := \arg \min_{x \in X} \|x - a_j\|_2$

2. Select a semi-inner product $\langle \Theta_1 \cdot, \Theta_1 \cdot \rangle_2$, $\Theta_1 \in \mathbb{R}^{l' \times m}$ $l' \geq l$, define

$$P^{so} = A\Omega_1 (\Theta_1 A\Omega_1)^+ \Theta_1,$$

or equiv $P^{so} a_j := \arg \min_{x \in X} \|\Theta_1 (x - a_j)\|_2$
Properties of the approximations

Definitions and some of the results taken from [Demmel et al., 2019].

**Definition 1**
[low-rank approximation] A matrix $A_k$ satisfying $\|A - A_k\|_2 \leq \gamma \sigma_{k+1}(A)$ for some $\gamma \geq 1$ will be said to be a $(k, \gamma)$ low-rank approximation of $A$.

**Definition 2**
[spectrum preserving] If $A_k$ satisfies

$$\sigma_j(A) \geq \sigma_j(A_k) \geq \gamma^{-1} \sigma_j(A)$$

for $j \leq k$ and some $\gamma \geq 1$, it is a $(k, \gamma)$ spectrum preserving.

**Definition 3**
[kernel approximation] If $A_k$ satisfies

$$\sigma_{k+j}(A) \leq \sigma_j(A - A_k) \leq \gamma \sigma_{k+j}(A)$$

for $1 \leq j \leq n - k$ and some $\gamma \geq 1$, it is a $(k, \gamma)$ kernel approximation of $A$. 

Deterministic rank-k matrix approximation

Given $A \in \mathbb{R}^{m \times n}$, $\Theta = \begin{pmatrix} \Theta_1 \\ \Theta_2 \end{pmatrix} \in \mathbb{R}^{m \times m}$, $\Omega = \begin{pmatrix} \Omega_1 & \Omega_2 \end{pmatrix} \in \mathbb{R}^{n \times n}$, $\Theta, \Omega$ invertible, $\Theta_1 \in \mathbb{R}^{l' \times m}$, $\Omega_1 \in \mathbb{R}^{n \times l}$, $k \leq l \leq l'$.

$$\Theta A \Omega = \bar{A} = \begin{pmatrix} \bar{A}_{11} & \bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{22} \end{pmatrix}$$

$$= \begin{pmatrix} I \\ \bar{A}_{21} \bar{A}_{11}^+ I \end{pmatrix} \begin{pmatrix} \bar{A}_{11} & \bar{A}_{12} \\ & S(\bar{A}_{11}) \end{pmatrix} = \Theta \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \begin{pmatrix} R_{11} & R_{12} \\ & R_{22} \end{pmatrix},$$

where $\bar{A}_{11} \in \mathbb{R}^{l',l}$, $\bar{A}_{11}^+ \bar{A}_{11} = I$, $S(\bar{A}_{11}) = \bar{A}_{22} - \bar{A}_{21} \bar{A}_{11}^+ \bar{A}_{12}$.

- Generalized LU computes the approximation

$$A_k = \Theta^{-1} \begin{pmatrix} I \\ \bar{A}_{21} \bar{A}_{11}^+ \end{pmatrix} (\bar{A}_{11} \bar{A}_{12}) \Omega^{-1}$$

- QR computes the approximation

$$A_k = Q_1 \begin{pmatrix} R_{11} & R_{12} \end{pmatrix} V^{-1} = Q_1 Q_1^T A,$$ where $Q_1$ is orth basis for $(A \Omega_1)$
Unified perspective: generalized LU factorization

Given $\Theta_1$, $A$, $\Omega_1$, $Q_1$ orth. basis of $(A\Omega_1)$, $k = l = l'$, rank-k approximation,

$$A_k = A\Omega_1 (\Theta_1 A\Omega_1)^{-1} \Theta_1 A$$

<table>
<thead>
<tr>
<th>Deterministic algorithms</th>
<th>Randomized algorithms*</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Omega_1$ column permutation and ...</td>
<td>$\Omega_1$ random matrix and ...</td>
</tr>
<tr>
<td>QR with column selection</td>
<td>Randomized QR</td>
</tr>
<tr>
<td>(a.k.a. strong rank revealing QR)</td>
<td>(a.k.a. randomized SVD)</td>
</tr>
<tr>
<td>$\Theta_1 = Q_1^T$, $A_k = Q_1 Q_1^T A$</td>
<td>$\Theta_1 = Q_1^T$, $A_k = Q_1 Q_1^T A$</td>
</tr>
<tr>
<td>$</td>
<td></td>
</tr>
<tr>
<td>LU with column/row selection</td>
<td>Randomized LU with row selection</td>
</tr>
<tr>
<td>(a.k.a. rank revealing LU)</td>
<td>(a.k.a. randomized SVD via Row extraction)</td>
</tr>
<tr>
<td>$\Theta_1$ row permutation s.t. $\Theta_1 Q_1 = \begin{pmatrix} \bar{Q}<em>{11} \ \bar{Q}</em>{21} \end{pmatrix}$</td>
<td>$\Theta_1$ row permutation s.t. $\Theta_1 Q_1 = \begin{pmatrix} \bar{Q}<em>{11} \ \bar{Q}</em>{21} \end{pmatrix}$</td>
</tr>
<tr>
<td>$</td>
<td></td>
</tr>
</tbody>
</table>

Deterministic algorithms will be discussed in a future lecture.
Given $\Theta_1, A, \Omega_1, Q_1$ orth. basis of $(A\Omega_1)$, $k \leq l \leq l'$, rank-k approximation,

$$A_k = [\Theta_1^+(I - (\Theta_1 A\Omega_1)(\Theta_1 A\Omega_1)^+) + (A\Omega_1)(\Theta_1 A\Omega_1)^+]\Theta_1 A], \quad (7)$$

where $\Theta_1$ and $(\Theta_1 A\Omega_1)$ are of dimensions $l' \times m$ and $l' \times l$ respectively.

**Remark** Given that only $\Theta_1$ and $\Omega_1$ are required for computing $A_k$, $\Theta$ and $\Omega$ are used only for the analysis, $\Theta_2$ and $\Omega_2$ are chosen to be the orthogonal of $\Theta_1$ and $\Omega_1$ respectively.
Properties of projection based approximations

Proposition 4 (Proposition 4.3 from [Demmel et al., 2019])

Let \( A \in \mathbb{R}^{m \times n} \) matrix with SVD \( A = U \Sigma V^T \). Set \([Q, R] = QR(A \Omega)\), where \( \Omega \in \mathbb{R}^{n \times n} \) matrix, \( \Omega = (\Omega_1, \Omega_2) \), \( \Omega_1 \) is full column rank and \( \Omega_2 \) is the orthogonal of \( \Omega_1 \). Then the singular values of \( Q_1 Q_1^T A - A \) are identical to those of matrix \( R_{22} \in \mathbb{R}^{(m-l) \times (n-l)} \). Moreover,

\[
\| R_{22} \|_F^2 \leq \| \Sigma_{1,2} \|_F^2 + \| \Sigma_{1,2}(V^T \Omega)_{21}(V^T \Omega)_{11}^+ \|_F^2
\]

\[
\sigma_j(A) \geq \sigma_j(Q_1 Q_1^T A) \geq \sigma_j(A \Omega_1) \sigma_{\min}(\Omega_1^+) \quad \text{for } j \leq k
\]

as well as for any given \( j \leq \min(m, n) - k \), there is an orthogonal \( n \times (n - j) \) matrix \( \tilde{V} \) independent of \( \Omega \) such that

\[
\sigma_j^2(R_{22}) \leq \sigma_{j+k}^2(A) + \| \Sigma_{j,2}(\tilde{V}^T \Omega)_{21}(\tilde{V}^T \Omega)_{11}^+ \|_2^2
\]

with \((\tilde{V}^T \Omega)_{11} \in \mathbb{R}^{k \times l}\), and \( \Sigma_{j,2} := \text{diag}(\sigma_{k+j}(A), \ldots, \sigma_n(A), 0, \ldots, 0) \), and \( \Sigma_{j,2} \in \mathbb{R}^{(m-k) \times (n-k)} \), where \( \text{diag} \) denotes the diagonal matrix.
Plan

Randomization for least-squares problem

Low rank matrix approximation

Randomized algorithms for low rank approximation
Randomized algorithms - main idea

- Construct a low dimensional subspace that captures the action of $A$.
- Restrict $A$ to the subspace and compute a standard QR or SVD factorization.

** Obtained as follows:**

1. Compute an approximate basis for the range of $A$ ($m \times n$) find $Q_1$ ($m \times k$) with orthonormal columns and approximate $A$ by the projection of its columns onto the space spanned by $Q_1$:

\[
A \approx Q_1 Q_1^T A
\]

2. Use $Q_1$ to compute a standard factorization of $A$

Typical randomized SVD

1. Compute an approximate basis for the range of $A \in \mathbb{R}^{m \times n}$
   Sample $\Omega_1 \in \mathbb{R}^{n \times l}$, $l = p + k$, with independent mean-zero, unit-variance Gaussian entries.
   Compute $Y = A\Omega_1$, $Y \in \mathbb{R}^{m \times l}$ expected to span column space of $A$.
   □ Cost of multiplying $A\Omega_1$: $2mnl$ flops

2. With $Q_1$ being orthonormal basis of $Y$, approximate $A$ as:

   $$\tilde{A}_k = Q_1 Q_1^T A = \mathcal{P}^o A$$

   □ Cost of multiplying $Q_1^T A$: $2mnl$ flops

Typical randomized SVD

Algorithm

**Input:** matrix \( A \in \mathbb{R}^{m \times n} \), desired rank \( k \), \( l = p + k \).

1. Sample an \( n \times l \) test matrix \( \Omega_1 \) with independent mean-zero, unit-variance Gaussian entries.
2. Compute \( Y = (AA^T)^{q}A\Omega_1 \) /* \( Y \) is expected to span the column space of \( A \) */
3. Construct \( Q_1 \in \mathbb{R}^{m \times l} \) with columns forming an orthonormal basis for the range of \( Y \).
4. Compute \( B = Q_1^T A, \ B \in \mathbb{R}^{l \times n} \)
5. Compute the rank-\( k \) truncated SVD of \( B \) as \( \hat{U}\Sigma V^T \), \( \hat{U} \in \mathbb{R}^{l \times k} \), \( V^T \in \mathbb{R}^{k \times n} \)

**Return** the approximation \( \tilde{A}_k = Q_1 \hat{U} \cdot \Sigma \cdot V^T \)
Randomized SVD \((q = 0)\)

The best approximation is when \(Q_1\) equals the first \(k + p\) left singular vectors of \(A\). Given \(A = U\Sigma V^T\),

\[
Q_1 Q_1^T A = U(1 : m, 1 : k + p)\Sigma(1 : k + p, 1 : k + p)(V(1 : n, 1 : k + p)
\]

\[
\|A - Q_1 Q_1^T A\|_2 = \sigma_{k+p+1}
\]

**Theorem 1.1** from Halko et al. If \(\Omega_1\) is chosen to be i.i.d. \(N(0, 1)\), \(k, p \geq 2, q = 1\), then the expectation with respect to the random matrix \(\Omega_1\) is

\[
\mathbb{E}(\|A - Q_1 Q_1^T A\|_2) \leq \left(1 + \frac{4\sqrt{k + p}}{p - 1} \sqrt{\min(m, n)}\right) \sigma_{k+1}(A)
\]

and the probability that the error satisfies

\[
\|A - Q_1 Q_1^T A\|_2 \leq \left(1 + 11\sqrt{k + p} \cdot \sqrt{\min(m, n)}\right) \sigma_{k+1}(A)
\]

is at least \(1 - 6/p^p\).

For \(p = 6\), the probability becomes \(.99\).
Theorem 10.6, Halko et al. Average spectral norm. Under the same hypotheses as Theorem 1.1 from Halko et al.,

\[
\mathbb{E}(\|A - Q_1Q_1^T A\|_2) \leq \left( 1 + \sqrt{\frac{k}{p - 1}} \right) \sigma_{k+1}(A) + \frac{e \sqrt{k + p}}{p} \left( \sum_{j=k+1}^{n} \sigma_j^2(A) \right)^{1/2}
\]

- Fast decay of singular values:
  If \( \left( \sum_{j>k} \sigma_j^2(A) \right)^{1/2} \approx \sigma_{k+1} \) then the approximation should be accurate.
- Slow decay of singular values:
  If \( \left( \sum_{j>k} \sigma_j^2(A) \right)^{1/2} \approx \sqrt{n - k \sigma_{k+1}} \) and \( n \) large, then the approximation might not be accurate.

Source: G. Martinsson’s talk
Power iteration $q \geq 1$

The matrix $(AA^T)^q A$ has a faster decay in its singular values:
- has the same left singular vectors as $A$
- its singular values are:

$$\sigma_j((AA^T)^q A) = (\sigma_j(A))^{2q+1}$$
Cost of randomized truncated SVD

- Randomized SVD requires $2q + 1$ passes over the matrix.
- The last 4 steps of the algorithm cost:
  1. Compute $Y = (AA^T)^q A \Omega_1$: $2(2q + 1) \cdot \text{nnz}(A) \cdot (k + p)$
  2. Compute QR of $Y$: $2m(k + p)^2$
  3. Compute $B = Q_1^T A$: $2\text{nnz}(A) \cdot (k + p)$
  4. Compute SVD of $B$: $O(n(k + p)^2)$
- If $\text{nnz}(A)/m \geq k + p$ and $q = 1$, then (2) and (4) dominate (3).
- To be faster than deterministic approaches, the cost of (2) and (4) need to be reduced.
Fast Johnson-Lindenstrauss transform

Find sparse or structured $\Omega_1$ such that computing $A\Omega_1$ is cheap, e.g. a subsampled random Hadamard transform (SRHT).

Given $n = 2^p$, $l < n$, the SRHT ensemble embedding $\mathbb{R}^n$ into $\mathbb{R}^l$ is defined as

$$\Omega_1 = \sqrt{\frac{n}{l}} \cdot P \cdot H \cdot D,$$

(10)

- $D \in \mathbb{R}^{n \times n}$ is diagonal matrix of uniformly random signs, random variables uniformly distributed on $\pm 1$
- $H \in \mathbb{R}^{n \times n}$ is the normalized Walsh-Hadamard transform
- $P \in \mathbb{R}^{l \times n}$ formed by subset of $l$ rows of the identity, chosen uniformly at random (draws $l$ rows at random from $HD$).

References: Sarlos’06, Ailon and Chazelle’06, Liberty, Rokhlin, Tygert and Woolfe’06.
Definition of Normalized Walsh–Hadamard Matrix
For given $n = 2^p$, $H_n \in \mathbb{R}^{n \times n}$ is the non-normalized Walsh-Hadamard transform defined recursively as,

$$H_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad H_n = \begin{pmatrix} H_{n/2} & H_{n/2} \\ H_{n/2} & -H_{n/2} \end{pmatrix}. \quad (11)$$

The normalized Walsh-Hadamard transform is $H = n^{-1/2}H_n$.

Cost of matrix vector multiplication (Theorem 2.1 in [Ailon and Liberty, 2008]):
For $x \in \mathbb{R}^n$ and $\Omega_1 \in \mathbb{R}^{l \times n}$, computing $\Omega_1x$ costs $2n \log_2(l + 1)$ flops.
Results from image processing (from Halko et al)

- A matrix $A$ of size $9025 \times 9025$ arising from a diffusion geometry approach.
- $A$ is a graph Laplacian on the manifold of $3 \times 3$ patches.
- $95 \times 95$ pixel grayscale image, intensity of each pixel is an integer $\leq 4095$.
- Vector $x^{(i)} \in \mathbb{R}^9$ gives the intensities of the pixels in a $3 \times 3$ neighborhood of pixel $i$.
- $W$ reflects similarities between patches, $\sigma = 50$ reflects the level of sensitivity,
  $$w_{ij} = \exp\left\{-\|x^{(i)} - x^{(j)}\|^2/\sigma^2\right\},$$
- Sparsify $W$, compute dominant eigenvectors of $A = D^{-1/2}WD^{-1/2}$. 
Experimental results (from Halko et al)

- Approximation error: $||A - Q_1 Q_1^T A||_2$
- Estimated eigenvalues for $k = 100$
Definition 5
A \((k, \epsilon, \delta)\) oblivious subspace embedding (OSE) from \(\mathbb{R}^n\) to \(\mathbb{R}^l\) is a distribution \(\Omega_1 \sim D\) over \(l \times n\) matrices. It satisfies with probability \(1 - \delta\)

\[
1 - \epsilon \leq \sigma_{\min}^2(\Omega_1 Q) \leq \sigma_{\max}^2(\Omega_1 Q) \leq 1 + \epsilon
\]

for any given orthogonal \(n \times k\) matrix \(Q\). We will assume \(l \geq k\) and \(\epsilon < 1/6\).

Definition 6
\(\Omega_1 \in \mathbb{R}^{l \times n}\) is \((\epsilon, \delta, n)\) multiplication approximating, if for any \(A, B\) having \(n\) rows, it satisfies with probability \(1 - \delta\),

\[
\|A^T \Omega_1^T \Omega_1 B - A^T B\|_F \leq \epsilon \|A\|_F \|B\|_F. \tag{12}
\]
Properties of SRHT ensembles

Additional property of the SRHT ensemble from Lemma 4.8 of [Boutsidis and Gittens, 2013].

**Lemma 7**

Let $\Omega_1$ be drawn from an SRHT of dimension $l \times n$. Then for $m \times n$ matrix $A$ with rank $\rho$, with probability $1 - 2\delta$,

$$\|A\Omega_1^T\|_2^2 \leq 5\|A\|_2^2 + \frac{\log(\rho/\delta)}{l}(\|A\|_F + \sqrt{8\log(n/\delta)}\|A\|_2^2)^2$$

Oblivious embeddings: Let $\Omega_1 \in \mathbb{R}^{l \times n}$ be drawn from SRHT ensembles. With $l = 4\epsilon^{-1}k(1 + 2\sqrt{\ln(3/\delta)})^2(1 + \sqrt{8\ln(3n/\delta)})^2$, $\Omega_1$ is a $(k, \sqrt{\epsilon}, 3\delta)$ OSE (Lemma 4.1 from [Boutsidis and Gittens, 2013]). It satisfies the multiplication property with $(\epsilon/k, \delta, n)$ (Lemma 4.11 from [Boutsidis and Gittens, 2013]).
Subspace embeddings

Lemma 5.4 from [Demmel et al., 2019], an extension of Lemma 4.1 of [Boutsidis and Gittens, 2013].

**Lemma 8**

Let $\Omega_1$ be an $l \times n$ matrix that is a $(k, \epsilon, \delta)$ OSE from $\mathbb{R}^n$ to $\mathbb{R}^l$, and $Q$ be an $(n \times k)$ orthogonal matrix. Provided $\epsilon < 1/6$, then with probability $1 - \delta$ both of the following hold,

$$
\|(\Omega_1 Q)^+ - (\Omega_1 Q)^T\|_2^2 \leq 3\epsilon
$$

(13)

$$
\|\Omega_1\|_2^2 = O\left(\frac{n}{k}\right),
$$

(14)

where in the second of these we require the additional assumption $\delta > 2e^{-k/5}$. 

Corollary 9 (Corollary 5.16 in [Demmel et al., 2019])

Let \( \Omega_1 \in \mathbb{R}^{n \times l} \) be drawn from an SHRT ensemble,
\( l \geq 4\epsilon^{-1}k(1 + 2\sqrt{\ln(3/\delta)})^2(1 + \sqrt{8 \ln(3n/\delta)})^2 \), \( \Omega_1 \), and for simplicity assume \( l \geq \log(n/\delta) \log(\rho/\delta) \). Then with probability \( 1 - 2\delta \)

\[
\sigma_j^2(R_{22}) \leq O(1)\sigma_{k+j}^2(A) + O\left(\frac{\log(\rho/\delta)}{l}\right)(\sigma_{k+j}^2(A) + \ldots \sigma_n^2(A)), \tag{15}
\]

for \( 1 \leq j \leq \min(m, n) - k \) with probability \( 1 - 3\delta \) for a particular \( j \). We also have upper and lower bounds on the largest singular values, as for \( 1 \leq j \leq k \),

\[
\sigma_j(A) \geq \sigma_j(Q_1Q_1^T A) = \Omega(\sqrt{\frac{k}{n}})\sigma_j(A), \tag{16}
\]

holds with probability \( 1 - 2 \max(\delta, e^{-k/5}) \).
Details of proof of eq (15)

Begin by using Proposition 4 and Lemma 8,

\[ \sigma^2_j(R_{22}) \leq \|\Sigma_{j,2}\|_2^2 + \|\Sigma_{j,2}(\tilde{V}^T\Omega)_{21}(\tilde{V}^T\Omega)_{11}^+\|_2^2 \leq \|\Sigma_{j,2}\|_2^2 + 2\|\Sigma_{j,2}(\tilde{V}^T\Omega)_{21}\|_2^2, \]

with probability $1 - \delta$. Next apply Lemma 7 to the second term to get

\[ \sigma^2_j(R_{22}) = O \left(1 + \frac{\log(\rho/\delta) \log(n/\delta)}{l}\right) \|\Sigma_{j,2}\|_2^2 + O \left(\frac{\log(\rho/\delta)}{l}\right) \|\Sigma_{j,2}\|_F^2 \]

\[ = O(1)\|\Sigma_{j,2}\|_2^2 + O \left(\frac{\log(\rho/\delta)}{l}\right) \|\Sigma_{j,2}\|_F^2 \] \hspace{1cm} (17)

\[ = O(1)\sigma^2_{k+j}(A) + O\left(\frac{\log(\rho/\delta)}{l}\right)(\sigma^2_{k+j}(A) + \ldots \sigma^2_n(A)), \] \hspace{1cm} (18)

where $\rho$ is the rank of $A$, with probability $1 - 2\delta$. 


Consider \( \Theta_1 \in \mathbb{R}^{l' \times m}, \Omega_1 \in \mathbb{R}^{n \times l} \) are Subsampled Randomized Hadamard Transforms (SRHT), \( l' > l \).

Compute \( A_k \) through generalized LU as in equation (7) costs \( O(mn \log_2 l' + ml') \) flops,

\[
A_k = [\Theta_1^+(I - (\Theta_1 A \Omega_1)(\Theta_1 A \Omega_1)^+) + (A \Omega_1)(\Theta_1 A \Omega_1)^+] [\Theta_1 A].
\]

**Theorem 10 (Theorem 5.9 from [Demmel et al., 2019])**

Let \( \Theta_1 \in \mathbb{R}^{l' \times m} \) and \( \Omega_1 \in \mathbb{R}^{n \times l} \) be drawn from SRHT ensembles,

\[
l = 4\epsilon^{-1} k (1 + 2\sqrt{\ln(3/\delta)})^2 (1 + \sqrt{8\ln(3n/\delta)})^2,
\]

\[
l' = 4\epsilon^{-1} l (1 + 2\sqrt{\ln(3/\delta)})^2 (1 + \sqrt{8\ln(3m/\delta)})^2.
\]

With probability \( 1 - 5\delta \), the randomized GLU approximation \( A_k \) satisfies

\[
\|A - A_k\|_2^2 = O(1)\sigma_{k+1}^2(A) + O(\frac{\log(n/\delta)}{l} + \frac{\log(m/\delta)}{l'}) (\sigma_{k+1}^2(A) + \ldots \sigma_n^2(A))
\]

\[
\sigma_j^2(A - A_k) \leq O(1)\sigma_{k+j}^2 + O(\frac{\log(\rho/\delta)}{l} + \frac{\log(\rho/\delta)}{l'}) (\sigma_{k+j}^2(A) + \ldots \sigma_n^2(A)).
\]
Fast dimension reduction using rademacher series on dual bch codes.

Improved matrix algorithms via the subsampled randomized hadamard transform.

An improved analysis and unified perspective on deterministic and randomized low rank matrix approximations.

Eckart, C. and Young, G. (1936).
The approximation of one matrix by another of lower rank.
*Psychometrika*, 1:211–218.

Relative perturbation techniques for singular value problems.

Sketching as a tool for numerical linear algebra.
Results used in the proofs

  Let $A = [a_1 | \ldots | a_n]$ be a column partitioning of an $m \times n$ matrix with $m \geq n$. If $A_r = [a_1 | \ldots | a_r]$, then for $r = 1 : n - 1$

  \[
  \sigma_1(A_{r+1}) \geq \sigma_1(A_r) \geq \sigma_2(A_{r+1}) \geq \ldots \geq \sigma_r(A_{r+1}) \geq \sigma_r(A_r) \geq \sigma_{r+1}(A_{r+1}).
  \]

- Given $n \times n$ matrix $B$ and $n \times k$ matrix $C$, then ([Eisenstat and Ipsen, 1995], p. 1977)

  \[
  \sigma_{\min}(B)\sigma_j(C) \leq \sigma_j(BC) \leq \sigma_{\max}(B)\sigma_j(C), j = 1, \ldots, k.
  \]