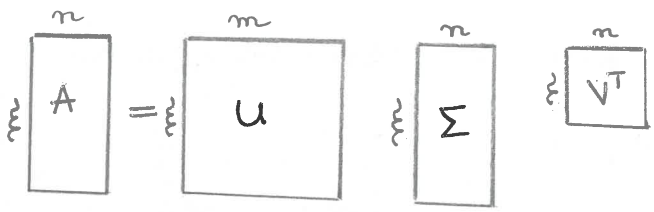
Divey Anand 2022aa05030 | Dibyajyoti Sarkar 2022aa05005 | Medha Jain 2022aa05083 | Shiffin N 2022aa05055

Q5 We need to send a 1000×1000 matrix of numbers across a channel, and would like to minimize the total amount of data sent on the channel for reasons having to do with both the possibility of data corruption and the time taken to send the data. Can you think of a way of minimizing the amount of data to be sent across the channel, so that we can represent the most important information in the matrix?

We know that any matrix $A \in \mathbb{R}^{m \times n}$ of rank $r \in [0, \min(m,n)]$ can always be factored into a Singular Value Decomposition (SVD) as follows:



U

$$\begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ u_1 & u_2 & \vdots & u_m \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

Σ

$$\begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_m \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix}$$

V^T

$$\begin{bmatrix} \cdots & v_1^T & \cdots \\ \cdots & v_2^T & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & v_n^T & \cdots \end{bmatrix}$$

Where,

- U is a $m \times m$ orthogonal matrix of column vectors u_i $(1 \geq i \geq m)$. where U essentially contains information about the column space of A and is also called the left-singular vectors.
- V is a $n \times n$ orthogonal matrix of column vectors v_i $(1 \geq i \geq n)$. Where V essentially contains information about the row space of A and is also called the right-singular vectors.
- Σ is a $m \times n$ matrix with $\Sigma_{ii} = \sigma_i \geq 0$ and $\Sigma_{ij} = 0, i \neq j$ where Σ essentially contains information about how important the columns of U & V are and is also called the singular values.
  The diagonal entries of Σ are ordered as $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0$.
  Σ has the same size as A. Meaning:
  - If $m > n$ then Σ has diagonal structure upto row n & then consists of 0s from $n+1$ to m
  - If $m < n$ then Σ has diagonal structure upto ~~row~~ column m & then consists of 0s from $m+1$ to n

Divey Anand 2022aa05030 | Dibyajyoti Sarkar 2022aa05005 | Medha Jain 2022aa05083 | Shiffin N 2022aa05055

⊛ The columns of U are hierarchically arranged such that column u_1 is more important than u_2 and so on. The rows of V are hierachically arranged such that row v_1 is more important than v_2 and so on. And their importance is encoded in the singular values σ.

We know the computing the full SVD of a large $m \times n$ matrix can be quite taxing. So, instead we will now demonstrate how SVD allows us to represent matrix A as a sum of simpler (low-rank) matrices A_i, which lends itself to a matrix approximation scheme that is cheaper than the full SVD.

We will now try to demonstrate the full SVD as a sum of rank 1 matrices A_i

$$A_{m \times n} = U \Sigma V^T = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ u_1 & u_2 & \vdots & u_n & \vdots & u_m \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \cdots & v_1^T & \cdots \\ \cdots & v_2^T & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & v_n^T & \cdots \end{bmatrix}$$

(where $m \gg n$)

$$= \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_n u_n v_n^T + 0$$

$$= \hat{U} \hat{\Sigma} V^T$$

⊛ Since Σ is a diagonal matrix when we multiply $U\Sigma$ column u_1 essentially gets scaled by σ_1, column u_2 by σ_2 and so on; and similarly, when we multiply $U \Sigma V^T$ the first column $\sigma_1 u_1$ only multiplies the v_1^T column, $\sigma_2 u_2$ column only multiplies the v_2^T column and so on

⊛ Eventhough the U matrix has m columns, there are only n non-zero singular values in the Σ matrix. So everything after the first n columns in U & V becomes 0.

Essentially what this means is that we can select just the n columns of U i.e. $\hat{U}$, the first $n \times n$ block in Σ i.e. $\hat{\Sigma}$ and the $n \times n$ matrix V^T and write that as $\hat{U} \hat{\Sigma} V^T$ and that is exactly the same as $A_{m \times n}$

Divey Anand 2022aa05030 | Dibyajyoti Sarkar 2022aa05005 | Medha Jain 2022aa05083 | Shiffin N 2022aa05055

Now that we have represented matrix A as a sum of rank 1 matrices A_i. We can intuitively see that —

- The best rank 1 approximation of A is $\sigma_1 u_1 v_1^T$
- The best rank 2 approximation of A is $\sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T$ and so on (and this is what SVD essentially means)

What we will now do is truncate out approximation of matrix A at rank k.

What this means is if we have a lot of small singular values σ_i $(k+1 \geq i \geq n)$ are negligibly small and most of the information about matrix A is captured in the k singular values and singular vectors. We can keep $\sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \dots + \sigma_k u_k v_k^T$ i.e. the first k columns of U and V^T, the first $k \times k$ submatrix of Σ and ignore the rest.

$$A_{m \times n} = U \Sigma V^T = \hat{U} \hat{\Sigma} V^T \approx \hat{U} \hat{\Sigma} \hat{V}^T$$

A formal definition of this type of approximation of A can be found in the Eckart-Young Theorem.

where a matrix $A \in R^{m \times n}$ of rank r and a matrix $B \in R^{m \times n}$ of rank k for any $k \leq r$ with $\hat{A}(k) = \sum_{i=1}^{k} \sigma_i u_i v_i^T$ it holds that —

$$\hat{A}(k) = \operatorname{argmin}_{rk(B)=k} = \|A - B\|_2, \; \|A - \hat{A}(k)\|_2 = \hat{U} \hat{\Sigma} \hat{V}^T$$

The Eckart-Young theorem implies that we can use SVD to reduce a rank r matrix A to a rank k matrix $\hat{A}$ in a principled, optimal (in the spectral norm sense) manner.

Therefore, in conclusion, by using the Eckart-Young Theorem we can approximate the most important information of matrix A by a rank k matrix (ie the first k columns of U and V, the first $k \times k$ submatrix of Σ) as a form of lossy compression.