Hajar Zaid
CUNY Graduate Center
Singular Value Decomposition

The Gram Matrix, Cosine Similarity

Gram Matrix: The gram matrix is defined as $G = X^\top X$.

Suppose we are given the following matrix: $\begin{bmatrix} 2 & 3 \\ 1 & 1 \end{bmatrix}$

$$G = \begin{bmatrix} 2 & 1 \\ 3 & 1 \end{bmatrix}\begin{bmatrix} 2 & 3 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 5 & 7 \\ 7 & 10 \end{bmatrix}$$

Each entry, G_{ij}, encodes the *dot product* between rows and columns. G_{11} is the dot product between the first row and first column, G_{12} is the dot product between the first row and second column, G_{21} is the dot product between the second row and first column, and G_{22} is the dot product between the second row and second column.

Let's take a closer look at the first entry of the Gram Matrix.

$$G_{11} = \begin{bmatrix} 2 & 1 \end{bmatrix}\begin{bmatrix} 2 \\ 1 \end{bmatrix} = 2^2 + 1^2 = 5$$

Just looking at the scalar quanitity here, the amount of alignment between these two vectors isn't immediately clear. What does a dot product of 5 mean? To make this more interpretable we *normalize* the quanitity by dividing by the magnitude of each vector.

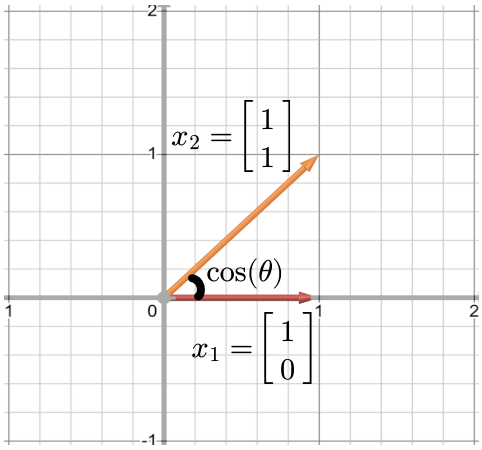
The magnitude of $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$ is simply the length of this vector. We can compute the euclidean distance using the ℓ_2 norm.

$$\frac{x_1^\top x_2}{\|x_1\|_2\|x_2\|_2}\frac{5}{\sqrt{5}\cdot\sqrt{5}} = \frac{5}{\sqrt{25}} = \frac{5}{5} = 1$$

Therefore, the *normalized* dot product of the first row and first column is 1. This is normalization of vector alignment is called ***cosine similarity.***

Cosine Similarity: The cosine similarity of two vectors x_1 and x_2 is defined as

$$\cos(\theta) = \frac{x_1^\top x_2}{\|x_1\|_2\|x_2\|_2}$$



Why $\cos(\theta)$?

We know that the hypotenuse is$\sqrt{2}$ just using pythagorean theorem.

The cosine angle between these two vectors is:

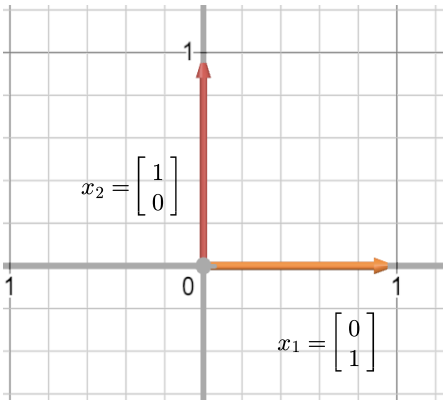
$$\cos(\theta) = \frac{adj}{hyp} = \frac{1}{\sqrt{2}}$$

Equivalently, from our cosine similarity formula above:

$$\cos(\theta) = \frac{x_1 \cdot x_2}{\|x_1\|_2 \|x_2\|_2} = \frac{1 \cdot 1 + 1 \cdot 0}{\sqrt{1^2 + 1^2} \cdot \sqrt{1^2 + 0^2}} = \frac{1}{\sqrt{2}}$$

Cosine similiarity values range from -1 to 1 where a value of 1 signifies perfect alignment and value of -1 signifies oppositely oriented vectors (180 degrees apart). A cosine similarity of 0 means the vectors are orthogonal to each other.

Example of Orthogonal Vectors:



$$\cos(\theta) = \frac{1 \cdot 0 + 0 \cdot 1}{\sqrt{1^2 + 0^2} + \sqrt{0^2 + 1^2}} = 0 \ \leftsquigarrow \ orthogonal$$

In summary, the Gram matrix is a matrix where each entry represents the dot product between pairs of vectors. Normalizing these dot products transforms them into cosine values, which reflect the angles between the vectors. Thus, a normalized Gram matrix encodes the angular relationships between the vectors.

Orthogonal and Orthonormal Matrices

Orthogonal Matrices are matrices where rows and columns are orthogonal (perpendicular) to one another. *Orthonormal Matrices* are matrices where rows and columns are orthogonal to one other AND are normalized to unit length (the ℓ_2 norm of each row and column is exactly equal to one).

All orthonormal matrices are orthogonal but not all orthogonal matrices are orthonormal.

Example of an Orthonormal Matrix

The identity matrix is always orthonormal because the rows and columns of the identity matrix are the standard basis vectors, which are orthogonal to each other and have a length of 1.

Example of a matrix that is orthogonal but NOT orthonormal

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$$

$\begin{bmatrix} 2 & 0 \end{bmatrix}\begin{bmatrix} 0 \\ 3 \end{bmatrix} = 0 \leftarrow$ dot product of 0 means orthogonal rows/columns.

However, the first and second row are not unit length and therefore this is not an orthonormal matrix.

Suppose we multiply the following orthonormal matrix by the vector defined below:

$$A = \begin{bmatrix} \sqrt{2}/2 & -\sqrt{2}/2 \\ \sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix}, x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

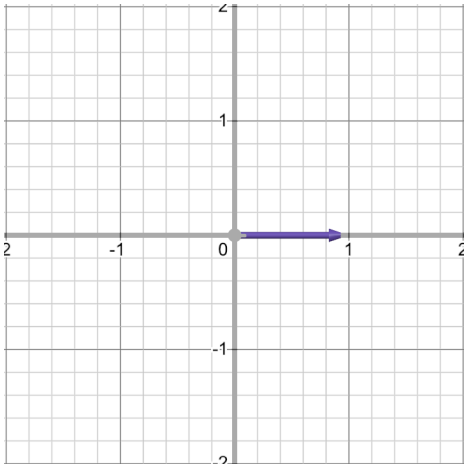


Figure 1: The original vector $x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

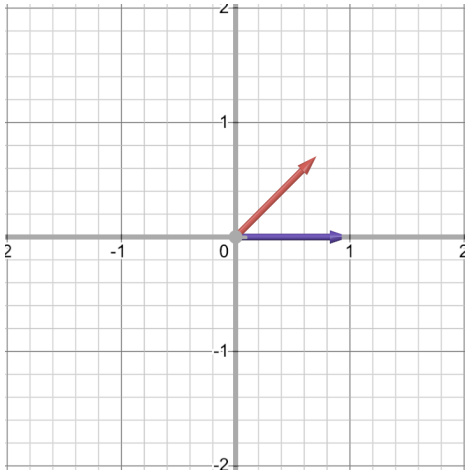


Figure 2: Apply the tranformation once; $x^{(1)} = Ax$.

$$x^{(1)} = \begin{bmatrix} \sqrt{2}/2 & -\sqrt{2}/2 \\ \sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \sqrt{2}/2 \\ \sqrt{2}/2 \end{bmatrix}$$

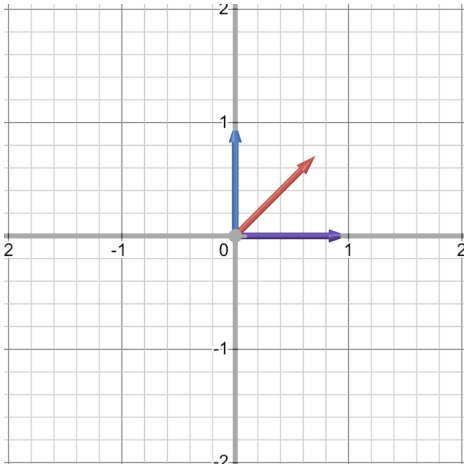


Figure 3: Applying the tranformation twice; $x^{(2)} = Ax^{(1)}$.

$$x^{(2)} = \begin{bmatrix} \sqrt{2}/2 & -\sqrt{2}/2 \\ \sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix}\begin{bmatrix} \sqrt{2}/2 \\ \sqrt{2}/2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

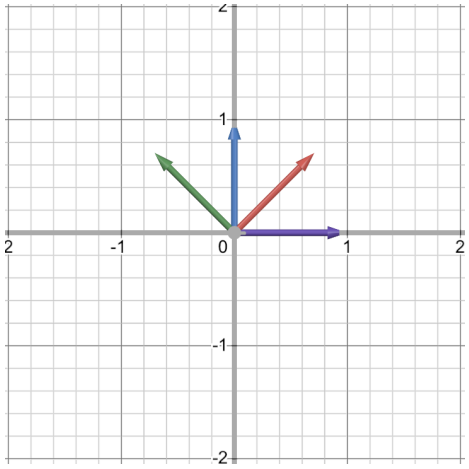


Figure 4: Applying the tranformation three times; $x^{(3)} = Ax^{(2)}$.

$$x^{(3)} = \begin{bmatrix} \sqrt{2}/2 & -\sqrt{2}/2 \\ \sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} -\sqrt{2}/2 \\ \sqrt{2}/2 \end{bmatrix}$$

Orthogonal Matrix:

Geometrically: An orthogonal matrix represents a rigid transformation (either a rotation or a reflection) that preserves the angles and distances between vectors.
Preserves lengths: The magnitude of vectors remains unchanged.
Preserves angles: The angle between vectors is maintained.
The rows and columns of the matrix are orthogonal to each other, but the vectors may not necessarily be unit vectors (they can have any magnitude).
Orthogonal matrices perform transformations like rotations and reflections, but the vectors involved in the transformation are not required to be of unit length.

Orthonormal Matrix:

Geometrically: An orthonormal matrix is a special case of an orthogonal matrix where each row and each column is not only orthogonal but also a unit vector (magnitude = 1).
Effect on Vectors:
Preserves lengths: The magnitude of vectors remains unchanged.
Preserves angles: The angle between vectors is maintained.
The rows and columns are both orthogonal and unit vectors (magnitude = 1).
wOrthonormal matrices represent pure rotations (no scaling or reflection), with the vectors forming an orthonormal basis. The transformation involves only unit-length vectors.

Mathematical Properties of both Orthonormal and Orthogonal Matrices:

- $A^\top A = I,\ AA^\top = I$
- Both orthogonal and orthonormal matrices must be square.
  - rectangular matrices where $B \in \mathbb{R}^{m\times n}$ where $m \geq n$ and $B^\top B = I_n$ are called *semi-orthogonal matrices, orthogonal row matrices,* or *orthogonal column matrices* depending on whether rows or columns form an orthonormal set.

Singular Value Decomposition (SVD) is a factorization technique that decomposes any matrix $A \in \mathbb{R}^{m\times n}$ into three simpler matrices.

$$A = U\Sigma V^\top$$

U is an $m \times m$ orthogonal matrix whose columns are called *left singular vectors* of A.
Σ is an $m \times n$ diagonal matrix containing the *singular values* of A.
$V^\top$is an $n \times n$ orthogonal matrix whose rows are called *right singular vectors* of A.

The singular values of A are ordered in decreasing order. That is

$$\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \ldots$$

$$A = \underbrace{\begin{bmatrix} | & | & & | \\ u_1 & u_2 & \ldots & u_m \\ | & | & & | \end{bmatrix}}_{m\times m} \underbrace{\begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_m \\ & & & 0 \end{bmatrix}}_{m\times n} \underbrace{\begin{bmatrix} — & v_1 & — \\ — & v_2 & — \\ — & v_3 & — \\ & \vdots & \\ — & v_n & — \end{bmatrix}}_{n\times n}$$

Note: The number of non-zero singular values is exactly equal to the rank of matrix A. $r = \min(m, n)$. In the decomposition above, we assume $r = \min(m, n) = m$. A trail of one or more zeros in the diagonal signifies that our matrix is in **full SVD** form.

Reduced SVD

If A has rank r, we can write a compact singular value decomposition:

$$A = U_r \Sigma_r V_r^\top$$

U_r is an $m \times r$ orthogonal matrix whose columns are the first r *left singular vectors* of A.
Σ_r is an $r \times r$ diagonal matrix containing only the non-zero *singular values* of A.
$V_r^\top$ is an $r \times n$ orthogonal matrix whose rows correspond to the first r *right singular vectors* of A.

$$A = \underbrace{\begin{bmatrix} | & | & & | \\ u_1 & u_2 & \dots & u_r \\ | & | & & | \end{bmatrix}}_{m \times r} \underbrace{\begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_r \end{bmatrix}}_{r \times r} \underbrace{\begin{bmatrix} — & v_1 & — \\ — & v_2 & — \\ — & v_3 & — \\ & \vdots & \\ — & v_n & — \end{bmatrix}}_{r \times n}$$

The singular value decompostion is related to eigenvectors and eigenvalues of the matrix A. The columns of U are the eigenvectors of $AA^\top$ and the columns of V are the eigenvectors of $A^\top A$. In other words, U encodes eigenvectors associated with how rows interact with columns, while V encodes the eigenvectors associated with how columns interact with rows (*see Gram Matrix notes above*).

An intuitive way to think about the singular value decomposition is by representing a matrix tranformation in terms the effect it has on the unit sphere.

Fact: The image of the unit sphere under any $m \times n$ matrix is a hyperellipse.

$V^\top$: Rotation (or change of basis) in the input space.
Σ: Scaling of the axes (stretching the unit sphere into an ellipse)
U: Rotation (or change of basis) in the output space and projection onto the new subspace.

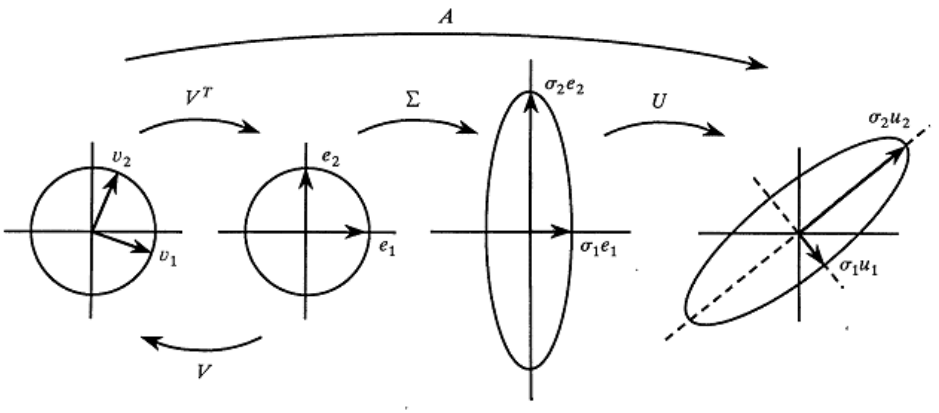


Figure 5: Left to Right: (1) V transforms basis vectors e_1 and e_2. Since $V^\top V = I$ (orthogonal) it preserves relative angles and distances of the axes. (2) Σ scales the axes by the singular values of $\sigma_1, \dots \sigma_n$, transforming the unit circle into an ellipse (3) U applies a final rotation to the ellipse, aligning it with the output basis.

In the image above, we can determine that A has full rank because the transformation preserves the two-dimensional structure; both the unit circle and the transformed ellipse remain in $\mathbb{R}^2$. If A were rank-deficient, it

would collapse some directions to lower dimensions, mapping a hypersphere to a lower-dimensional hyperellipse.

For example, if a matrix $A \in \mathbb{R}^{m \times n}$ maps n-dimensional vectors to m-dimensional vectors, then the n-dimensional unit sphere (or unit circle in 2D) would be mapped to an m-dimensional hyperellipse. If A has rank $r < \min(m, n)$, then the image of the sphere would be constrained to an r-dimensional subspace, effectively reducing its dimension.

Fact: Every matrix has a singular value decomposition (SVD), but it is not necessarily unique. The singular values in the decomposition are unique, but the matrices U and V (from $A = U\Sigma V^\top$) are not unique. If A has repeated singular values, the corresponding columns of U and V can be arbitrarily chosen within the subspaces spanned by the eigenvectors associated with those singular values, leading to multiple valid SVDs.

The SVD can be manually computed by computing the eigenvalues and eigenvectors of $A^T A$ and AA^T, extracting the singular values from the square roots of the nonzero eigenvalues, and constructing $U, \Sigma,$ and V accordingly.

SVD Calculation Steps:

Compute $A^\top A$:
Take the transpose of A and multiply it by A.

Find Eigenvalues of $A^\top A$:
Solve the characteristic equation $\det(A^T A - \lambda I) = 0$ *to find the eigenvalues* $\lambda_1, \lambda_2, \ldots$

Compute Singular Values:
The singular values σ_i are the square roots of the non-negative eigenvalues of $A^\top A$.

Find Eigenvectors of $A^\top A$:
For each eigenvalue λ_i, solve $(A^\top A - \lambda_i I)v_i = 0$ to find the eigenvectors v_i, which form the matrix V.

Compute the Left Singular Vectors (Matrix U):
For each eigenvector v_i of $A^\top A$, compute the corresponding left singular vector u_i by solving $Av_i = \sigma_i u_i$.
The matrix U consists of the left singular vectors u_i. Alternatively you can repeat the first 4 steps on $AA^\top$.

Construct the Matrix Σ:
Form a diagonal matrix Σ with singular values σ_i on the diagonal.

Final SVD:
The SVD of A is $A = U\Sigma V^\top$, where U is the matrix of left singular vectors, Σ is the diagonal matrix of singular values, and $V^\top$ is the transpose of the matrix of right singular vectors.

Computing the Singular Value Decomposition (SVD) directly is computationally expensive because it requires forming and diagonalizing $A^\top A$ or $AA^\top$, which involves $\mathcal{O}(n^3)$ operations for an $n \times n$ matrix. This approach is not only slow for large matrices but also numerically unstable due to potential loss of precision.

Low Rank Approximation with SVD
We care about Singular Value Decomposition because it provides a powerful way to analyze matrices by decomposing them into simpler components. This decomposition can help with tasks like solving linear systems, matrix inversion, and finding low-rank approximations, which are essential for dimensionality reduction, data

compression, and noise reduction.

Dimensionality reduction is performed by truncating the smaller singular values. Truncating the smallest singular values reduces the rank of the matrix, keeping only the most significant features (those corresponding to the largest singular values). This process results in a low-rank approximation of the original matrix, which captures the most important information while reducing the data's complexity and noise.

Example of *truncation:*

Given some matrix $A \in \mathbb{R}^{4\times 4}$, we write out its SVD as follows:

$$\begin{bmatrix} | & | & | & | \\ u_1 & u_2 & u_3 & u_4 \\ | & | & | & | \end{bmatrix} \begin{bmatrix} 100.5 & & & \\ & 95.2 & & \\ & & 76.8 & \\ & & & 2.4 \end{bmatrix} \begin{bmatrix} — & v_1 & — \\ — & v_2 & — \\ — & v_3 & — \\ — & v_4 & — \end{bmatrix}$$

Since the last singular value is smaller in comparison to the other singular values, we delete the last one as the eigenvectors associated with the singular value $\sigma = 2.4$ captures the least information about the data.

$$\begin{bmatrix} | & | & | \\ u_1 & u_2 & u_3 \\ | & | & | \end{bmatrix} \begin{bmatrix} 100.5 & & \\ & 95.2 & \\ & & 76.8 \end{bmatrix} \begin{bmatrix} — & v_1 & — \\ — & v_2 & — \\ — & v_3 & — \end{bmatrix}$$
