The Geometry of Information: Singular Value Decomposition and Data Compression

Authors: Anwen Hao, Jed Yao, Andy Chen

Abstract

While eigendecomposition is a powerful tool for analyzing square matrices, it is limited by requirements of diagonalizability. Singular Value Decomposition (SVD) offers a universal generalization applicable to any $m \times n$ matrix, factoring it into a rotation, a scaling, and another rotation. This paper explores the theoretical underpinnings of SVD through two distinct lenses: an algebraic derivation utilizing the Spectral Theorem, and a topological approach that characterizes singular values as variational extrema on the unit sphere. We illustrate these concepts through a worked example of the shear matrix, uncovering a direct connection between its singular values and the golden ratio via the Fibonacci matrix. Furthermore, we provide a rigorous proof of the Eckart-Young Theorem for the spectral norm. Finally, we demonstrate the practical utility of these results through the lens of image compression, along with some results used in modern algorithms for computing the SVD of large matrices.

Introduction

The central triumph of introductory linear algebra is the Spectral Theorem, which allows us to decompose symmetric matrices into an orthogonal basis of eigenvectors. However, the real world rarely presents us with data that is perfectly square or symmetric. In fields ranging from quantum mechanics to machine learning, we are often confronted with rectangular $m \times n$ matrices that defy standard eigendecomposition. How can we uncover the fundamental geometry of a transformation when the domain and codomain have different dimensions?

The answer lies in Singular Value Decomposition (SVD). SVD generalizes the Spectral Theorem to any matrix, regardless of shape or symmetry. This universality makes SVD indispensable for modern data science, particularly in dimensionality reduction and data compression, where we seek to approximate massive datasets with lower-rank matrices that preserve the most significant features. The utility of Eckart-Young has increased with the massive data sets used for A.I. Beyond standard image compression, where the rank-k approximation is used on the matrix of the image, the Eckart-Young-Mirsky Theorem is also a foundational result for Principal Component Analysis (PCA). PCA follows the same idea of dimension reduction on data to speed up A.I. training by reducing complexity of data.

This paper explores the theoretical foundations and practical applications of Singular Value Decomposition. We begin in **Section 1** by deriving SVD through two distinct approaches: a standard algebraic proof relying on the spectral properties of A^TA, and a topological or geometric characterization that treats singular values as extrema on the unit sphere. To illustrate these concepts, we provide a worked example of the shear matrix, uncovering an interesting connection between its singular values, the golden ratio, and Fibonacci numbers.

Building on this framework, **Section 2** proves the Eckart-Young Theorem for the spectral norm, rigorously establishing why the truncation of singular values yields the optimal low-rank approximation of a matrix. Finally, in **Section 3**, we introduce results used in the practical computation of SVD, and we apply these results to the problem of image compression, demonstrating how the abstract geometry of high-dimensional spaces directly translates to efficient data storage and compression. The goal of this section is to explain fundamental results that allow for the efficient computation of SVD, rather than dissecting these methods themselves in great detail. The hope is to provide enough information for a strong baseline understanding of the practical computation and application of SVD.

1 Singular Value Decomposition

1.1 Algebraic Characterization of SVD

We begin by walking through a proof for Singular Value Decomposition. This proof is highly algebraic and is based on the lecture notes for the class and relies on the key result from the Spectral Theorem. We will later use a more topological or geometric approach to derive the same result.

Theorem 1 (Singular Value Decomposition)**.** *Any $m \times n$ matrix A can be written as*

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

where U is an $m \times m$ orthogonal matrix, Σ is an $m \times n$ diagonal matrix with singular values along its diagonal, and V^T is an $n \times n$ orthogonal matrix. Geometrically, this means the transformation A can be decomposed into a rotation, a stretching, and another rotation.

Before proving this main result, we need to rely on a couple lemmas/propositions. The first observation is that the eigenvalues of A^TA are real and nonnegative:

Let $A \in \mathbb{R}^{m\times n}$, then the eigenvalues of A^TA are real and nonnegative.

Proof: We begin by briefly noting that A^TA is always symmetric:

$$(A^TA)^T = A^T(A^T)^T = A^TA.$$

Since A^TA is symmetric, by the Spectral Theorem, it must have real eigenvalues. To further prove that these eigenvalues are nonnegative, we calculate the square of the norm of $A\mathbf{v}$ for an arbitrary eigenvector $\mathbf{v}$ of A^TA.

$$\|A\mathbf{v}\|^2 = A\mathbf{v} \cdot A\mathbf{v} = \mathbf{v}^TA^TA\mathbf{v}$$

Since $\mathbf{v}$ is an eigenvector of A^TA, $A^TA\mathbf{v} = \lambda\mathbf{v}$, where λ is an eigenvalue for A^TA.

$$\mathbf{v}^TA^TA\mathbf{v} = \mathbf{v}^T\lambda\mathbf{v} = \lambda\mathbf{v} \cdot \mathbf{v} = \lambda\|\mathbf{v}\|^2$$

We can see that any arbitrary eigenvalue for A^TA can be expressed as

$$\lambda = \frac{\|A\mathbf{v}\|^2}{\|\mathbf{v}\|^2}$$

$\mathbf{v} \neq \mathbf{0}$ since $\mathbf{v}$ is an eigenvector, so this expression is always defined and greater than or equal to 0.

Definition 1 (singular values)**.** The *singular values* of $A \in \mathbb{R}^{m\times n}$ are the square roots of the eigenvalues of $A^T A \in \mathbb{R}^{n\times n}$.

Singular values are typically listed with their algebraic multiplicities (possible duplicates) and in decreasing order:

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$$

We will now prove a major result that will be essential in our final construction of SVD:

Lemma 1. *Let $A \in \mathbb{R}^{m\times n}$, then there exists an orthonormal basis $\mathbf{v}_1, \cdots, \mathbf{v}_n \in \mathbb{R}^n$ such that*

1. *$A\mathbf{v}_1, \cdots, A\mathbf{v}_n$ are orthogonal,*
2. *The lengths of $A\mathbf{v}_i$ are the singular values σ_i of A.*

Proof: As proven above, $A^T A$ has nonnegative eigenvalues. List these eigenvalues in decreasing order: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Since $A^T A$ is symmetric, by the Spectral Theorem, there exists an orthonormal eigenbasis. Construct this eigenbasis by finding the vectors $\mathbf{v}_1, \cdots, \mathbf{v}_n$ that correspond to these eigenvalues.

1. To prove that $A\mathbf{v}_i \perp A\mathbf{v}_j$, we need to show that $A\mathbf{v}_i \cdot A\mathbf{v}_j = 0$.

$$A\mathbf{v}_i \cdot A\mathbf{v}_j = (A\mathbf{v}_j)^T A\mathbf{v}_i = \mathbf{v}_j^T A^T A\mathbf{v}_i = \lambda_i \mathbf{v}_i \cdot \mathbf{v}_j = 0 \text{ if } i \neq j$$

2. To find the length of $A\mathbf{v}_i$, we find the norm of this vector. In a previous proof, we showed that

$$\|A\mathbf{v}_i\|^2 = \lambda_i \|\mathbf{v}_i\|^2$$

Taking the square root of both sides,

$$\|A\mathbf{v}_i\| = \sqrt{\lambda_i}\|\mathbf{v}_i\| = \sigma_i$$

since $\|\mathbf{v}_i\| = 1$.

Now, we make a statement about the relationship between singular values and rank:

Lemma 2. *If rank$(A) = r$, then $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ and $\sigma_{r+1} = \cdots = \sigma_n = 0$.*

Proof: Assume that there are some singular values greater than 0 and others equal to 0: $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_s > 0$ and $\sigma_{s+1} = \cdots = \sigma_n = 0$. Here, s is just an arbitrary number for the moment. We will show that this number s is equal to r, which is the rank of A.

We construct an orthonormal basis $\mathbf{v}_1, \cdots, \mathbf{v}_n$ according to the previous proof. In the previous proof, we showed that $\|A\mathbf{v}_i\| = \sigma_i$. Since $\sigma_i = 0$ for $i = s+1, \cdots, n$, it follows that $A\mathbf{v}_i$ must be the zero vector because the norm of a vector is 0 if and only if the vector itself is the zero vector. We now show that the remaining vectors $A\mathbf{v}_1, \cdots, A\mathbf{v}_s$ form the basis of the image of A; if we can show that, then s is the dimension of the image of A, which is by definition the rank.

In the previous proof, we showed that $A\mathbf{v}_1, \cdots, A\mathbf{v}_s$ are orthogonal to each other, which means that they are linearly independent. To show that these vectors span the image of A, take any vector in the image of A, $A\mathbf{v}$, where $\mathbf{v} \in \mathbb{R}^n$ is some vector in the domain. Since $\mathbf{v} \in \mathbb{R}^n$, it can be written as a linear combination of $\mathbf{v}_1, \cdots, \mathbf{v}_n$ because $\mathbf{v}_1, \cdots, \mathbf{v}_n$ is a basis for $\mathbb{R}^n$:

$$\mathbf{v} = c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$$

Applying A to both sides,

$$A\mathbf{v} = A(c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n) = c_1A\mathbf{v}_1 + \cdots + c_sA\mathbf{v}_s + c_{s+1}A\mathbf{v}_{s+1} + \cdots + c_nA\mathbf{v}_n$$

$A\mathbf{v}_{s+1} = \cdots = A\mathbf{v}_n = \mathbf{0}$, so $c_{s+1}A\mathbf{v}_{s+1} + \cdots + c_nA\mathbf{v}_n = \mathbf{0}$, which leaves us with

$$A\mathbf{v} = c_1A\mathbf{v}_1 + \cdots + c_sA\mathbf{v}_s.$$

This means that any vector $A\mathbf{v} \in \text{im}(A)$ can be written as a linear combination of $A\mathbf{v}_1, \cdots, A\mathbf{v}_s$, which means these vectors span the image of A.

Since $A\mathbf{v}_1, \cdots, A\mathbf{v}_s$ are linearly independent and span $\text{im}(A)$, they form a basis for $\text{im}(A)$, which means $\dim(\text{im}(A)) = \text{rank}(A) = s = r$.

We are now ready to fully construct the proof for Singular Value Decomposition.

Proof: Construct an orthonormal eigenbasis $\mathbf{v}_1, \cdots, \mathbf{v}_n \in \mathbb{R}^n$ as we did in a previous proof. Put these vectors as column vectors into an $n \times n$ matrix V:

$$V = \begin{bmatrix} | & & | \\ \mathbf{v}_1 & \cdots & \mathbf{v}_n \\ | & & | \end{bmatrix}$$

This matrix is orthogonal because the columns are orthonormal. Based on the previous results, we know that we can write $A\mathbf{v}_i$ in the following way:

$$A\mathbf{v}_i = \begin{cases} \sigma_i\mathbf{u}_i & \text{if } 1 \leq i \leq r \\ \mathbf{0} & \text{if } i > r \end{cases}$$

In other words, $A\mathbf{v}_i$ can be expressed as a scaling of a unit vector pointing in the same direction as $A\mathbf{v}_i$. This unit vector will thus be defined as $\mathbf{u}_i = \frac{A\mathbf{v_i}}{\sigma_i}$ for $i = 1, \cdots, r$. $\mathbf{u}_1, \cdots \mathbf{u}_r$ is a linearly independent (more precisely, orthonormal) set in $\mathbb{R}^m$, so we can extend this set into a basis for the entire space of $\mathbb{R}^m$. Put these vectors into an $m \times m$ matrix U:

$$U = \begin{bmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_m \\ | & & | \end{bmatrix}$$

Again, U is orthogonal because the columns are orthonormal. We now construct the full statement of SVD.

We begin by computing AV:

$$AV = A\begin{bmatrix} | & & | \\ \mathbf{v}_1 & \cdots & \mathbf{v}_n \\ | & & | \end{bmatrix} = \begin{bmatrix} | & & | \\ A\mathbf{v}_1 & \cdots & A\mathbf{v}_n \\ | & & | \end{bmatrix} = \begin{bmatrix} | & & | & | & & | \\ \sigma_1\mathbf{u}_1 & \cdots & \sigma_r\mathbf{u}_r & \mathbf{0} & \cdots & \mathbf{0} \\ | & & | & | & & | \end{bmatrix}$$

Now, we can pick out the singular values and put them in a diagonal matrix:

$$\begin{bmatrix} | & & | & | & & | \\ \sigma_1\mathbf{u}_1 & \cdots & \sigma_r\mathbf{u}_r & \mathbf{0} & \cdots & \mathbf{0} \\ | & & | & | & & | \end{bmatrix} = \begin{bmatrix} | & & | & | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_r & \mathbf{0} & \cdots & \mathbf{0} \\ | & & | & | & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & & 0 \\ & \ddots & & \\ & & \sigma_r & \\ 0 & & & 0 \end{bmatrix}$$

Now, the matrix $\begin{bmatrix} | & & | & | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_r & \mathbf{0} & \cdots & \mathbf{0} \\ | & & | & | & & | \end{bmatrix}$ is an $m \times m$ matrix. Since the diagonal entries after σ_r are all 0, we can replace the $\mathbf{0}$ column vectors by the extended basis vectors $\mathbf{u}_{r+1}, \cdots, \mathbf{u}_m$ and the product would not change.

$$\begin{bmatrix} | & & | & | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_r & \mathbf{0} & \cdots & \mathbf{0} \\ | & & | & | & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & & 0 \\ & \ddots & & \\ & & \sigma_r & \\ 0 & & & 0 \end{bmatrix} = \begin{bmatrix} | & & | & | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_r & \mathbf{u}_{r+1} & \cdots & \mathbf{u}_m \\ | & & | & | & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & & 0 \\ & \ddots & & \\ & & \sigma_r & \\ 0 & & & 0 \end{bmatrix} = U\Sigma$$

Recall that this entire expression is originally equal to AV:

$$AV = U\Sigma$$

V is an orthogonal matrix, so we can multiply on the right by V^T, finishing the proof:

$$AVV^T = U\Sigma V^T$$

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

1.2 Topological Characterization of SVD

While the previous section derives Singular Value Decomposition algebraically using the Spectral Theorem, on A^TA, we can essentially rediscover SVD geometrically by asking a simple yet fundamental question: *In which direction does the matrix A maximize the length of a vector?*

This variational or topological approach offers a distinct insight, which is that singular values are not merely hidden eigenvalues, but rather the extrema of the linear transformation on the unit sphere. This thought process of optimization will segue nicely into our discussion of Eckart-Young.

Here is a restatement of Singular Value Decomposition with this geometric twist:

Theorem 2 (Topological Characterization of SVD)**.** *Let $A \in \mathbb{R}^{m\times n}$, then there exist unit vectors $\mathbf{v}_1 \in \mathbb{R}^n$ and $\mathbf{u}_1 \in \mathbb{R}^m$ such that $A\mathbf{v}_1 = \sigma_1\mathbf{u}_1$, where $\sigma_1 = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$. Furthermore, for any unit vector $\mathbf{w} \in S$, where $S = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| = 1\}$, if $\mathbf{w} \perp \mathbf{v}_1$, then $A\mathbf{w} \perp A\mathbf{v}_1$.*

Before we proceed, we will clarify what we mean by compact set, directly quoting from the definition of compactness from [1].

Definition 2 (Compact Set)**.** "A subset X of $\mathbb{R}^n$ is compact if, for every sequence $\{\mathbf{v}_k\}$ with $\mathbf{v}_k$ for all k, there exists a subsequence $\{\mathbf{v}_{k_i}\}$ which converges to a limit $\mathbf{w} \in X$."

Proof: Let $S = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| = 1\}$ be the unit sphere. Since S is a compact set and $\|A\mathbf{x}\|$ is continuous, there must exist a vector $\mathbf{v}_1 \in S$ that maximizes $\|A\mathbf{x}\|$. This existence fact comes from the following theorem, quoted directly from Robert Friedman's Honors Math notes [1].

Theorem 3. *"If X is a nonempty compact subset of $\mathbb{R}^n$, then every continuous function $f : X \to \mathbb{R}$ satisfies the Extreme Value Theorem."*

The Extreme Value Theorem states that there must exist a maximum and minimum value for a continuous function on a closed and bounded interval. Applied to our case, the above theorem means precisely what we stated earlier, provided the unit sphere is compact: there must exist a vector $\mathbf{v}_1 \in S$ that maximizes $\|A\mathbf{x}\|$, where $\|A\mathbf{x}\|$ serves as the continuous function.

We show that the unit sphere is indeed compact. We have a sequence $\{\mathbf{v}_k\}$. Since $\mathbf{v}_k$ lies on the unit sphere, $\|\mathbf{v}_k\| = 1$. This means this sequence is bounded. Now, we rely on a fundamental result from analysis known as the Bolzano-Weierstrass Theorem.

Theorem 4 (Bolzano-Weierstrass)**.** *Every bounded sequence in $\mathbb{R}^n$ has a convergent subsequence.*

Since our sequence $\{\mathbf{v}_k\}$ is bounded, by this theorem, there must exist a subsequence $\{\mathbf{v}_{k_i}\}$ that converges to a limit $\mathbf{w} \in \mathbb{R}^n$.

But to satisfy the definition of compactness, this limit must lie on the unit sphere, so we need to show $\|\mathbf{w}\| = 1$.

Due to the continuity of the norm, as $\mathbf{v}_{k_i}$ approaches $\mathbf{w}$, the norm of $\mathbf{v}_{k_i}$ also approaches the norm of $\mathbf{w}$. Therefore,

$$\|\mathbf{w}\| = \lim_{i\to\infty} \|\mathbf{v}_{k_i}\| = 1$$

since all elements in the subsequence also lie on the unit sphere, so $\|\mathbf{v}_{k_i}\| = 1$. The limit that the subsequence converges to lies on the unit sphere, so the unit sphere is indeed a compact set.

We return to our original goal, which is to maximize $\|A\mathbf{x}\|$. Equivalently, we can maximize the function $f(\mathbf{x}) = \|A\mathbf{x}\|^2$ because both $\|A\mathbf{x}\|$ and $\|A\mathbf{x}\|^2$ are greater than or equal to 0, and $\|A\mathbf{x}\|^2$ yields cleaner algebra. So, our objective is now to find

$$\max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|^2$$

We can rewrite $\|A\mathbf{x}\|^2$ using the dot product:

$$f(\mathbf{x}) = \|A\mathbf{x}\|^2 = A\mathbf{x} \cdot A\mathbf{x} = \mathbf{x}^T A^T A\mathbf{x}$$

Another way to interpret this maximizing problem is to think that we are optimizing the function $f(\mathbf{x})$ *subject to a constraint*, namely $\|\mathbf{x}\| = 1$. We can rewrite this constraint as another function, $g(\mathbf{x})$:

$$\|\mathbf{x}\| = \mathbf{x}^T\mathbf{x} = 1$$

$$g(\mathbf{x}) = \mathbf{x}^T\mathbf{x} - 1 = 0$$

In multivariable calculus, when we are trying to optimize a multivariate function subject to a constraint, we use a tool called Lagrange multipliers.

The key observation that Lagrange multipliers rely on is the fact that the extrema occur precisely where the contour lines of the multivariate function are tangent to the constraint function. Another way to frame this is that the gradient vectors of f and the constraint g are parallel. Since the gradients are parallel, they are scalar multiples of each other:

$$\nabla f(\mathbf{x}) = \lambda \nabla g(\mathbf{x})$$

As of now, the λ represents some constant. We compute the gradients. While we may compute the gradients using brute force by computing the partial derivatives of each, the gradient describes the first order change in f, or in other words, the best linear approximation of f can be described using the gradient:

$$f(\mathbf{x}+\mathbf{h}) \approx f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot \mathbf{h}$$

So, if we expand out $f(\mathbf{x}+\mathbf{h})$ and keep only the linear term, we look at the term that multiplies $\mathbf{h}$, and that term will be the gradient.

We expand out $g(\mathbf{x}+\mathbf{h})$:

$$g(\mathbf{x}+\mathbf{h}) = (\mathbf{x}+\mathbf{h})^T(\mathbf{x}+\mathbf{h}) - 1 = \mathbf{x}^T\mathbf{x} + 2\mathbf{x}^T\mathbf{h} + \mathbf{h}^T\mathbf{h} - 1 =$$

$$= (\mathbf{x}^T\mathbf{x} - 1) + 2\mathbf{x}^T\mathbf{h} + \mathbf{h}^T\mathbf{h} = g(\mathbf{x}) + 2\mathbf{x}^T\mathbf{h} + \mathbf{h}^T\mathbf{h} = g(\mathbf{x}) + (2\mathbf{x}) \cdot \mathbf{h} + \mathbf{h}^T\mathbf{h}$$

We see that $2\mathbf{x}$ multiplies $\mathbf{h}$, so the gradient of $g(\mathbf{x})$ must be $2\mathbf{x}$:

$$\nabla g(\mathbf{x}) = 2\mathbf{x}$$

We now expand $f(\mathbf{x}+\mathbf{h})$. Let $M = A^TA$:

$$f(\mathbf{x}+\mathbf{h}) = (\mathbf{x}+\mathbf{h})^TM(\mathbf{x}+\mathbf{h}) = \mathbf{x}^TM\mathbf{x} + \mathbf{x}^TM\mathbf{h} + \mathbf{h}^TM\mathbf{x} + \mathbf{h}^TM\mathbf{h} = f(\mathbf{x}) + \mathbf{x}^TM\mathbf{h} + \mathbf{h}^TM\mathbf{x} + \mathbf{h}^TM\mathbf{h}$$

$\mathbf{h}^TM\mathbf{h}$ is the quadratic term, so we focus on the linear terms $\mathbf{x}^TM\mathbf{h} + \mathbf{h}^TM\mathbf{x}$. Note that $(\mathbf{x}^TM\mathbf{h})^T = \mathbf{x}^TM\mathbf{h}$ since it is a scalar:

$$(\mathbf{x}^TM\mathbf{h})^T = \mathbf{h}^TM^T\mathbf{x} = \mathbf{h}^TM\mathbf{x} = \mathbf{x}^TM\mathbf{h}$$

Above we used the fact that $M = A^TA$ is symmetric, so $M = M^T$. The linear terms combine to become:

$$\mathbf{x}^TM\mathbf{h} + \mathbf{h}^TM\mathbf{x} = 2\mathbf{h}^TM\mathbf{x} = 2M\mathbf{x} \cdot \mathbf{h}$$

The gradient of $f(\mathbf{x})$ is thus:

$$\nabla f(\mathbf{x}) = 2M\mathbf{x} = 2A^TA\mathbf{x}$$

Plugging these gradients into the Lagrange multiplier condition, we have:

$$2A^TA\mathbf{x} = \lambda(2\mathbf{x})$$

$$A^T A\mathbf{x} = \lambda\mathbf{x}$$

λ used to be some unknown constant, but the above statement shows that the maximizing vectors are precisely the eigenvectors of $A^T A$. This directly connects this geometric interpretation to the algebraic notion of eigenvectors.

Originally, we supposed $\mathbf{v}_1 \in S$ maximizes $f(\mathbf{x}) = \|A\mathbf{x}\|^2 = \mathbf{x}^T A^T A\mathbf{x}$. Then $\mathbf{v}_1$ is an eigenvector, $A^T A\mathbf{v}_1 = \lambda_1\mathbf{v}_1$. We find the maximum value $f(\mathbf{x})$ attains by plugging in $\mathbf{v}_1$:

$$\max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|^2 = \max_{\|\mathbf{x}\|=1} f(\mathbf{x}) = f(\mathbf{v}_1) = \|A\mathbf{v}_1\|^2 = \mathbf{v}_1^T A^T A\mathbf{v}_1 = \mathbf{v}_1^T \lambda_1\mathbf{v}_1 = \lambda_1(\mathbf{v}_1 \cdot \mathbf{v}_1) = \lambda_1\|\mathbf{v}_1\|^2 = \lambda_1$$

This allows us to find σ_1:

$$\sigma_1 = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\| = \sqrt{\lambda_1}$$

This eigenvalue must be the largest of the eigenvalues of A because of a theorem we know from the lecture notes, which is that if M is symmetric, its quadratic form can be expressed as

$$q(\mathbf{x}) = \sum_{i=1}^{n} \lambda_i c_i^2$$

This means

$$f(\mathbf{x}) = \mathbf{x}^T M\mathbf{x} = \sum_{i=1}^{n} \lambda_i c_i^2$$

since $f(\mathbf{x})$ is a quadratic form. The maximum the quadratic form can achieve is the largest eigenvalue:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \lambda_i c_i^2 \leq \sum_{i=1}^{n} \lambda_{\max} c_i^2 = \lambda_{\max} \sum_{i=1}^{n} c_i^2 = \lambda_{\max}$$

Above, $\sum_{i=1}^{n} c_i^2 = 1$ since $\|\mathbf{x}\| = 1$ and c_i represent the coordinates of $\mathbf{x}$.

We already showed above that the maximum $f(\mathbf{x})$ can attain is λ_1:

$$\max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|^2 = \max_{\|\mathbf{x}\|=1} f(\mathbf{x}) = f(\mathbf{v}_1) = \lambda_1$$

This means that λ_1 must be the maximum eigenvalue, $\lambda_{\max}$.

To summarize, we have shown the existence of a maximizing vector $\mathbf{v}_1$, and this vector must be an eigenvector. We have just now shown that the length of the resultant vector $A\mathbf{v}_1$ has length $\sigma_1 = \sqrt{\lambda_1}$. We can define the resultant vector as $A\mathbf{v}_1 = \sigma_1\mathbf{u}_1$.

The orthogonality claim $\mathbf{w} \perp \mathbf{v}_1 \Rightarrow A\mathbf{w} \perp A\mathbf{v}_1$ follows immediately:

$$A\mathbf{v}_1 \cdot A\mathbf{w} = \mathbf{v}_1^T A^T A\mathbf{w} = A^T A\mathbf{v}_1 \cdot \mathbf{w} = \lambda_1(\mathbf{v}_1 \cdot \mathbf{w}) = 0$$

This implies that A maps the orthogonal complement of $\mathbf{v}_1$ into the orthogonal complement of $\mathbf{u}_1$. We know what happens when A is applied to $\mathbf{v}_1$. As we have defined, $A\mathbf{v}_1 = \sigma_1\mathbf{u}_1$. Now, consider the vectors orthogonal to $\mathbf{v}_1$ in the domain like $\mathbf{w}$ (these vectors compose the orthogonal complement of $\mathbf{v}_1$). As we have just shown, for any vector $\mathbf{w} \perp \mathbf{v}_1$, $A\mathbf{w} \perp A\mathbf{v}_1$, or $A\mathbf{w} \perp \sigma_1\mathbf{u}_1 \Rightarrow A\mathbf{w} \perp \mathbf{u}_1$. Put simply, given $\mathbf{w} \perp \mathbf{v}_1$, applying A results in a vector that is perpendicular to $\mathbf{u}_1$.

We can complete the derivation for SVD by induction. Restrict the domain of A to the orthogonal complement of $\mathbf{v}_1$, $\text{span}(\mathbf{v}_1)^\perp$. There must be a vector $\mathbf{v}_2 \in \text{span}(\mathbf{v}_1)^\perp$ that maximizes $\|A\mathbf{x}\|$ for the same reason in the proof above: the norm is a continuous function, and the unit sphere is a compact set. The same Lagrange multiplier steps will yield $\mathbf{v}_2$ as an eigenvector, and $\sigma_2 = \sqrt{\lambda_2}$. Define $\mathbf{u}_2 = \frac{A\mathbf{v}_2}{\sigma_2}$.

Note that since we found $\mathbf{v}_2 \in \text{span}(\mathbf{v}_1)^\perp$, $\mathbf{v}_2$ is perpendicular to $\mathbf{v}_1$. We proved above that A maps the orthogonal complement of the maximizing vector to the orthogonal complement of the image of the maximizing vector. Therefore, $\mathbf{v}_1 \perp \mathbf{v}_2 \Rightarrow A\mathbf{v}_1 \perp A\mathbf{v}_2 \Rightarrow \mathbf{u}_1 \perp \mathbf{u}_2$.

We keep repeating this process, finding $\mathbf{v}_i, \mathbf{u}_i, \sigma_i$ until we run out of dimensions. This inductive procedure effectively stops when we have a singular value equal to 0, or when we run out of dimensions in the domain or codomain in the case of a rectangular matrix.

Let r be the number of nonzero singular values. We have two cases:

1. If $r < n$ (input basis for $\mathbb{R}^n$), we choose orthonormal vectors from the kernel of A to complete the basis for $\mathbb{R}^n$. This is because the singular values being zero effectively imply that A maps everything in the subspace to $\mathbf{0}$, which is by definition the kernel.

2. If $r < m$ (output basis for $\mathbb{R}^m$), we extend the orthonormal set we have to an entire basis of $\mathbb{R}^m$.

The expression $A\mathbf{v}_i = \sigma_i\mathbf{u}_i$ thus holds for all vectors, which yield the SVD:

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

The idea of varying over the unit sphere is in part inspired by Trefethen and Bau [2], though we do not use the notion of a hyperellipsoid. Using Lagrange multipliers is a result of our independent thinking and knowledge from multivariable calculus.

1.3 Worked Example

We perform an SVD manually on a small, toy example to show SVD in action. We will use the 2×2 shear matrix:

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

We observe that this matrix is not very friendly in the sense that it is neither symmetric nor diagonalizable, as we will show. It is precisely in situations like these where SVD provides a clean decomposition of the matrix into more friendly matrices, particularly orthogonal and diagonal matrices. However, at the same time, SVD is very labor-intensive, so performing SVD on larger matrices is largely done by software.

We first find the eigenvalues for this matrix.

$$\det(A - \lambda I) = \det(\begin{bmatrix} 1-\lambda & 1 \\ 0 & 1-\lambda \end{bmatrix}) = (1-\lambda)^2 = 0$$

We can see that this matrix has one eigenvalue $\lambda = 1$ with algebraic multiplicity 2. We find the corresponding eigenvector and eigenspace for this eigenvalue.

$$\ker(\begin{bmatrix} 1-1 & 1 \\ 0 & 1-1 \end{bmatrix}) = \ker(\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix})$$

(Fun Fact: This matrix is what we call a "nilpotent" matrix, or a matrix for which $A^k = 0$ for some number k) So, we have to solve the system

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \mathbf{0}$$

Our result is $y = 0$, but x can be any real number. Thus, our eigenspace is

$$E_\lambda = \text{span}(\begin{bmatrix} 1 \\ 0 \end{bmatrix})$$

This space has dimension 1, or geometric multiplicity 1. Thus, $\text{gemu}(\lambda) < \text{almu}(\lambda)$, so this shear matrix is not diagonalizable.

We now perform SVD on the shear matrix. We first need to find an orthonormal eigenbasis for A^TA.

$$A^TA = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

We find the eigenvalues for this matrix.

$$\det(\begin{bmatrix} 1-\lambda & 1 \\ 1 & 2-\lambda \end{bmatrix}) = (1-\lambda)(2-\lambda) - 1 = 2 - 3\lambda + \lambda^2 - 1 = \lambda^2 - 3\lambda + 1 = 0$$

The eigenvalues are thus $\lambda_1 = \frac{3+\sqrt{5}}{2}$, which happens to be the square of the golden ratio, ϕ^2, and $\lambda_2 = \frac{3-\sqrt{5}}{2}$, which happens to be ϕ^{-2}. This ordering is to ensure that the singular values are listed in decreasing order. Each eigenvalue has algebraic multiplicity 1. We find eigenvectors corresponding to these eigenvalues.

We find the eigenvector of λ_1.

$$\ker(A^TA - \phi^2 \cdot I) = \ker(\begin{bmatrix} 1-\phi^2 & 1 \\ 1 & 2-\phi^2 \end{bmatrix})$$

$\phi^2 = \phi + 1$, so the matrix simplifies to

$$\ker(\begin{bmatrix} -\phi & 1 \\ 1 & 1-\phi \end{bmatrix})$$

We have the equations

$$\begin{cases} -\phi x + y = 0 \\ x + (1-\phi)y = 0 \end{cases}$$

The first equation becomes $y = \phi x$, and the second equation becomes

$$x + y - \phi y = 0$$

$$x + \phi x - \phi^2 x = x + \phi x - (\phi + 1)x = x + \phi x - \phi x - x = 0$$

We see that the left side completely simplifies to 0, so the only constraint we have is $y = \phi x$. An eigenvector that satisfies this is $\mathbf{v}_1 = \begin{bmatrix} 1 \\ \phi \end{bmatrix}$.

We now find the eigenvector for λ_2.

$$\ker(A^T A - \phi^{-2} \cdot I) = \ker(\begin{bmatrix} 1-\phi^{-2} & 1 \\ 1 & 2-\phi^{-2} \end{bmatrix})$$

The top left expression simplifies to

$$1 - \phi^{-2} = 1 - \frac{1}{\phi^2} = \frac{\phi^2 - 1}{\phi^2} = \frac{\phi}{\phi^2} = \phi^{-1} = \phi - 1$$

The bottom right expression simplifies to

$$2 - \phi^{-2} = 1 + (1 - \phi^{-2}) = 1 + (\phi - 1) = \phi$$

Thus, our equations simplify elegantly into

$$\begin{cases} (\phi - 1)x + y = 0 \\ x + \phi y = 0 \end{cases}$$

We have $x = -\phi y$. The top equation simplifies into

$$\phi x - x + y = -\phi^2 y + \phi y + y = -(\phi + 1)y + \phi y + y = -\phi y - y + \phi y + y = 0$$

The left side completely simplifies to 0, so our only constraint is $x = -\phi y$. One eigenvector that satisfies this is $\mathbf{v}_2 = \begin{bmatrix} -\phi \\ 1 \end{bmatrix}$.

We note that the singular values are $\sigma_1 = \phi \geq \sigma_2 = \phi^{-1}$.

We are almost there with constructing an orthonormal eigenbasis because $\mathbf{v}_1 \perp \mathbf{v}_2$: $\mathbf{v}_1 \cdot \mathbf{v}_2 = -\phi + \phi = 0$. However, their norms are not one, so we need to make them unit length:

$$\mathbf{v}_1 = \frac{1}{\sqrt{1+\phi^2}} \begin{bmatrix} 1 \\ \phi \end{bmatrix}, \mathbf{v}_2 = \frac{1}{\sqrt{1+\phi^2}} \begin{bmatrix} -\phi \\ 1 \end{bmatrix}$$

We put these two vectors into our orthogonal V matrix.

$$V = \frac{1}{\sqrt{1+\phi^2}} \begin{bmatrix} 1 & -\phi \\ \phi & 1 \end{bmatrix}$$

We take the transpose of V, which happens to be the same matrix.

$$V^T = \frac{1}{\sqrt{1+\phi^2}} \begin{bmatrix} 1 & \phi \\ -\phi & 1 \end{bmatrix}$$

We now construct the U matrix using the definition of the $\mathbf{u}_i$ vectors as in our proof.

$$\mathbf{u}_1 = \frac{A\mathbf{v}_1}{\sigma_1} = \frac{1}{\phi} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \frac{1}{\sqrt{1+\phi^2}} \begin{bmatrix} 1 \\ \phi \end{bmatrix} = \frac{1}{\phi} \frac{1}{\sqrt{1+\phi^2}} \begin{bmatrix} 1+\phi \\ \phi \end{bmatrix} = \frac{1}{\phi} \frac{1}{\sqrt{1+\phi^2}} \begin{bmatrix} \phi^2 \\ \phi \end{bmatrix} = \frac{1}{\sqrt{1+\phi^2}} \begin{bmatrix} \phi \\ 1 \end{bmatrix}$$

$$\mathbf{u}_2 = \frac{A\mathbf{v}_2}{\sigma_2} = \frac{1}{\phi^{-1}} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \frac{1}{\sqrt{1+\phi^2}} \begin{bmatrix} -\phi \\ 1 \end{bmatrix} = \frac{1}{\phi^{-1}} \frac{1}{\sqrt{1+\phi^2}} \begin{bmatrix} 1-\phi \\ 1 \end{bmatrix} = \phi \frac{1}{\sqrt{1+\phi^2}} \begin{bmatrix} -\frac{1}{\phi} \\ 1 \end{bmatrix} = \frac{1}{\sqrt{1+\phi^2}} \begin{bmatrix} -1 \\ \phi \end{bmatrix}$$

We put these two vectors into the U matrix:

$$U = \frac{1}{\sqrt{1+\phi^2}} \begin{bmatrix} \phi & -1 \\ 1 & \phi \end{bmatrix}$$

The Σ matrix is simply the singular values on the diagonal:

$$\Sigma = \begin{bmatrix} \phi & 0 \\ 0 & \phi^{-1} \end{bmatrix}$$

Ultimately, our full Singular Value Decomposition of the shear matrix is

$$A = U\Sigma V^T = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = \frac{1}{\sqrt{1+\phi^2}} \begin{bmatrix} \phi & -1 \\ 1 & \phi \end{bmatrix} \begin{bmatrix} \phi & 0 \\ 0 & \phi^{-1} \end{bmatrix} \frac{1}{\sqrt{1+\phi^2}} \begin{bmatrix} 1 & \phi \\ -\phi & 1 \end{bmatrix}$$

The golden ratio permeates throughout this composition. This is actually not a surprise because this matrix has a connection to the Fibonacci matrix. Specifically, the Fibonacci matrix is defined as

$$F = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$$

It turns out this matrix generates the Fibonacci sequence. Consider the vector

$$\begin{bmatrix} f_{n-1} \\ f_n \end{bmatrix}$$

This vector contains the $(n-1)$th and nth Fibonacci number. Now, multiply this vector by the Fibonacci matrix:

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} f_{n-1} \\ f_n \end{bmatrix} = \begin{bmatrix} f_n \\ f_{n-1} + f_n \end{bmatrix} = \begin{bmatrix} f_n \\ f_{n+1} \end{bmatrix}$$

As we can see, multiplying by the Fibonacci matrix results in obtaining the next Fibonacci number. This is a recursive definition, but we can write an explicit formula:

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}^n \begin{bmatrix} f_0 \\ f_1 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}^n \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} f_n \\ f_{n+1} \end{bmatrix}$$

It turns out that A^TA that we calculated with the shear matrix is equal to the square of the Fibonacci matrix:

$$A^TA = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

$$F^2 = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

$$F^2 = A^TA$$

Instead of directly finding the eigenvalues of A^TA, we can find the eigenvalues for F and then square those eigenvalues because we know that if λ is an eigenvalue for an arbitrary matrix M, then λ^2 is an eigenvalue for M^2. This can be shown through a short proof. If λ is an eigenvalue for M, the following must be true:

$$M\mathbf{v} = \lambda\mathbf{v}$$

Apply M to both sides:

$$M^2\mathbf{v} = M(\lambda\mathbf{v})$$

If $\mathbf{v}$ is an eigenvector corresponding to eigenvalue λ, then all vectors in span($\mathbf{v}$) are eigenvectors corresponding to λ. Therefore, $\lambda\mathbf{v}$ is also an eigenvector corresponding to λ, so the following must be true:

$$M(\lambda\mathbf{v}) = \lambda(\lambda\mathbf{v}) = \lambda^2\mathbf{v}$$

Chaining this all together,

$$M^2\mathbf{v} = M(\lambda\mathbf{v}) = \lambda^2\mathbf{v}$$

Therefore, if we find the eigenvalues for F, then we know the eigenvalues of $F^2 = A^TA$ are precisely the squares of the eigenvalues of F.

We find the eigenvalues of F:

$$\det(F - \lambda I) = \det(\begin{bmatrix} -\lambda & 1 \\ 1 & 1-\lambda \end{bmatrix}) = \lambda^2 - \lambda - 1 = 0$$

The solutions to this characteristic equation are, by definition, the golden ratio and its negative reciprocal:

$$\phi = \frac{1+\sqrt{5}}{2}, -\phi^{-1} = \frac{1-\sqrt{5}}{2}$$

This means the eigenvalues for F^2 and thus A^TA are thus the squares of these numbers, which matches the eigenvalues we obtained through finding the eigenvalues for A^TA directly:

$$\lambda_1 = \phi^2, \lambda_2 = \phi^{-2}$$

As expected, the singular values are

$$\sigma_1 = \phi, \sigma_2 = \phi^{-1}$$

Ultimately, since A^TA is the square of the Fibonacci matrix, the appearance of the golden ratio is not accidental. The eigenvalues of A^TA are the squares of the eigenvalues of F, and square rooting those eigenvalues gives us back the golden ratio as the singular values.

1.4 Animation

As noted above, geometrically speaking, SVD can be interpreted as decomposing a linear transformation into a rotation, a scaling, and another rotation. Thus, SVD lends itself to some beautiful, visual animations. Find our animation for SVD here:

https://github.com/Anwen0507/singular-value-decomposition.

2 Eckart-Young Theorem

In this section, we state the relevant definitions needed to prove the Eckart-Young Theorem for the spectral norm, which is significant because it gives the best low-rank approximation of a matrix.

2.1 Rank-k Approximation

A rank-k approximation of a matrix A is any matrix B with $\text{rank}(B) \leq k$ that minimizes the reconstruction error:

$$\min_{\text{rank}(B)\leq k} \|A - B\|$$

This varies based on the exact norm function over a space.

2.2 Spectral Norm

The spectral norm (also called the operator 2-norm) of a matrix A is defined as

$$\|A\|_2 = \max_{\|\mathbf{x}\|_2=1} \|A\mathbf{x}\|_2,$$

where $\mathbf{x}$ is any unit vector in the domain of A. One important result is that the Spectral Norm is equal to the largest singular value of A. We start by applying SVD to the Euclidean norm of $A\mathbf{x}$:

$$\|A\mathbf{x}\|_2 = \|U\Sigma V^T\mathbf{x}\|_2 = \|\Sigma V^T\mathbf{x}\|_2.$$

Since U is orthogonal, $\|U\mathbf{x}\|_2 = \|\mathbf{x}\|_2$ for all $\mathbf{x}$. Moreover, because V^T is orthogonal, $\|V^T\mathbf{x}\|_2 = \|\mathbf{x}\|_2 = 1$.

We know that V is the change of basis matrix from the standard basis $\{\mathbf{e}_i\}$ to the orthonormal eigenbasis $\{\mathbf{v}_i\}$ of A^TA. Therefore, V is an orthogonal matrix, so $V^T = V^{-1}$, which implies that V^T is the change of basis matrix from $\{\mathbf{v}_i\}$ back to $\{\mathbf{e}_i\}$.

Any vector $\mathbf{x}$ can be written in terms of the basis $\{\mathbf{v}_i\}$ as

$$\mathbf{x} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n,$$

where the coefficients $c_1, \ldots, c_n$ satisfy $\|\mathbf{x}\|_2 = 1$.

Applying V^T, we get

$$V^T\mathbf{x} = c_1\mathbf{e}_1 + c_2\mathbf{e}_2 + \cdots + c_n\mathbf{e}_n.$$

Thus, the coefficients of $V^T\mathbf{x}$ in the standard basis $\{\mathbf{e}_i\}$ are exactly the coefficients of $\mathbf{x}$ in the orthonormal eigenbasis $\{\mathbf{v}_i\}$. We know that this change of basis simply expresses every unit vector in a different orthonormal basis.

Therefore, the maximum value of $\|\Sigma V^T\mathbf{x}\|_2$ occurs when $V^T\mathbf{x} = (1, 0, 0, \cdots)^T$, so that only the first (largest) singular value σ_1 contributes, giving

$$\max_{\|\mathbf{x}\|_2=1} \|\Sigma V^T\mathbf{x}\|_2 = \sigma_1$$

because the singular values are ordered in decreasing order.

Therefore,

$$\max_{\|\mathbf{x}\|_2=1} \|A\mathbf{x}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|\Sigma V^T\mathbf{x}\|_2 = \sigma_1,$$

2.3 Proof for Spectral Norm

We will prove the Eckart-Young theorem for the spectral norm. To do this, we need to show that the best rank k approximation for A is found by taking the first k singular values (the largest singular values) and setting the remaining values to 0. This proof follows a standard approach, slightly adapted from [3].

Theorem 5 (Eckart-Young-Mirsky Theorem for Spectral Norm)**.** *Let A be an $m \times n$ real matrix with rank r. Let its Singular Value Decomposition be given by $A = U\Sigma V^T$, with singular values ordered such that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(m,n)} \geq 0$.*

For any integer $1 \leq k < r$, let A_k be the truncated matrix defined by the first k singular values:

$$A_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

Then, for any matrix $B \in \mathbb{R}^{m\times n}$ with rank$(B) \leq k$:

$$\|A - A_k\|_2 = \min_{rank(B)\leq k} \|A - B\|_2$$

Furthermore, the minimal approximation error is determined by the $(k+1)$th singular value:

$$\|A - A_k\|_2 = \sigma_{k+1}$$

First, if we have $A_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ which is the matrix A truncated after the kth singular value, then the norm $||A - A_k||$ becomes

$$|| \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^T - \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T ||_2 = || \sum_{i=k+1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^T ||_2 = \sigma_{k+1}$$

because the spectral norm is equivalent to the largest singular value, and they are listed in decreasing order. Next, we have to show that for any other matrix B with rank$(B) \leq k$,

$$||A - B||_2 \geq \sigma_{k+1}$$

To do this, we first decompose B into two rank k matrices C and R via column row decomposition.

Let $B \in \mathbb{R}^{m\times n}$ be an arbitrary rank k approximation for A, and let $C \in \mathbb{R}^{m\times k}, R \in \mathbb{R}^{k\times n}$, where the k columns of C, $\mathbf{c}_i$, are the linearly independent columns of B.

Rewrite the columns of B the columns of C: $\mathbf{b}_i = a_{1i}\mathbf{c}_1 + a_{2i}\mathbf{c}_2 + \cdots + a_{ki}\mathbf{c}_k$ and let the columns of R, $R_i = [a_{1i}, a_{2i}, \cdots, a_{ki}]^T$. Then,

$$B = CR = \begin{bmatrix} | & & | \\ \mathbf{c}_1 & \cdots & \mathbf{c}_k \\ | & & | \end{bmatrix} \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{k1} & \alpha_{k2} & \cdots & \alpha_{kn} \end{bmatrix}$$

From this construction, rank$(C) =$ rank$(B) = k$.

Additionally, it can be shown that $\text{rank}(R) = k$ by looking at the row space of B.

If $B = CR$, then $\text{im}(B^T) \subseteq \text{im}(R^T)$ and hence $\text{rank}(B^T) \leq \text{rank}(R^T)$ Since for any matrix A, $\text{rank}(A) = \text{rank}(A^T)$,

$$\text{rank}(B) = k \leq \text{rank}(R).$$

Additionally, $R \in \mathbb{R}^{k \times n}$, so the row space of R can be at most dimension k. Hence,

$$\text{rank}(R^T) = \text{rank}(R) \leq k.$$

$$\text{rank}(B) = k \leq \text{rank}(R) \leq k$$

From this inequality, $\text{rank}(R) = k$.

We decompose B, $B = CR$, with $\text{rank}(C) = \text{rank}(R) = k$. The goal is to show the existence of a unit vector $\mathbf{x}$ in $\mathbb{R}^n$ such that $R\mathbf{x} = 0$, which would provide a lower bound for $||A - B_k||_2$ when applied to $\mathbf{w}$, and simultaneously cancel the B_k matrix entirely for simplicity.

To show the existence of such a vector, we first look at the dimension of the kernel of R. From the rank-nullity theorem on R,

$$n = k + \text{nullity}(R)$$

Rearranging, we get

$$\dim(\ker(R)) = n - k$$

Adding the dimensions of $\ker(R)$ and $\text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_{k+1})$, the span of the first $k+1$ right singular vectors, we get $\dim(\ker(R)) + \dim(V_{k+1}) = n+1$ because $\dim(\ker(R)) + \dim(V_{k+1}) = n+1$, $\ker(R) \cap \text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_{k+1}) \neq \{0\}$. This can be shown by contradiction.

Assume, for contradiction, that the intersection is trivial. Let $\{\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_{n-k}\}$ be a basis for $\ker(R)$. Then, we can show $\{\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_{n-k}, \mathbf{v}_1, \cdots, \mathbf{v}_{k+1}\}$ is a basis for the direct sum $\ker(R) \oplus \text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_{k+1})$.

To show linear independence, write any vector in the direct sum as a linear combination of the basis vectors and set to 0. Note that any $\mathbf{r} + \mathbf{v}$ can be written from the set because it contains a basis for each $\mathbf{r}$ and $\mathbf{v}$, and therefore the set spans the direct sum.

$$\mathbf{r} + \mathbf{v} = c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + \cdots + c_{n-k}\mathbf{a}_{n-k} + d_1\mathbf{v}_1 + \cdots + d_{k+1}\mathbf{v}_{k+1} = 0$$

Apply R to both sides, which cancels each term in the kernel of R:

$$R(d_1\mathbf{v}_1 + \cdots + d_{k+1}\mathbf{v}_{k+1}) = 0$$

Since we assumed the intersection of the kernel of R and the $\text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_{k+1})$ is trivial, then

$$d_1\mathbf{v}_1 + \cdots + d_{k+1}\mathbf{v}_{k+1} = 0$$

because these vectors are linearly independent, $d_1 = d_2 = \cdots = d_{k+1} = 0$.

Returning to the original equation, we can plug in the zeroes, and obtain

$$c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + \cdots + c_{n-k}\mathbf{a}_{n-k} = 0$$

and therefore $c_1, \cdots, c_{n-k} = 0$.

Because $\ker(R) \subseteq \mathbb{R}^n$ and $\text{span}(\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_{k+1}) \subset \mathbb{R}^n$, for any vector $\mathbf{r} \in \ker(R)$ and any $\mathbf{v} \in \text{span}(\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_{k+1})$, $\mathbf{r} + \mathbf{v} \in \mathbb{R}^n$ because $\mathbb{R}^n$ is closed under addition. Therefore, $\ker(R) \oplus \text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_{k+1}) \subseteq \mathbb{R}^n$. From the basis we constructed, $\dim(\mathbb{R}^n) \geq n+1$, which is a contradiction, meaning there must be a non-trivial intersection.

Taking a nonzero vector $\mathbf{w}$ in $\text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_{k+1})$ in the kernel of R (and therefore in the kernel of $B = CR$), we choose it to be of unit length and apply it to the norm:

$$\max_{\|\mathbf{x}\|_2=1} \|(A-B)\mathbf{x}\|_2 \geq ||(A-B)\mathbf{w}||_2 = ||A\mathbf{w}||_2$$

The inequality holds because the spectral norm is the maximum possible value for all vectors.

If $\mathbf{w} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_n$, we can write out the spectral norm $||A\mathbf{w}||_2^2 = a_1^2\sigma_1^2 + \cdots + a_{k+1}^2\sigma_{k+1}^2$.

Let $A = U\Sigma V^T$ be the singular value decomposition of A, where $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ and $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ are orthonormal bases of right and left singular vectors, respectively, and $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n)$ with $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$.

Let

$$\mathbf{w} = \sum_{i=1}^{n} a_i\mathbf{v}_i, \qquad \|\mathbf{w}\|_2 = 1.$$

Using $A\mathbf{v}_i = \sigma_i\mathbf{u}_i$, we obtain

$$A\mathbf{w} = \sum_{i=1}^{n} a_i\sigma_i\mathbf{u}_i.$$

We compute the squared norm:

$$\|A\mathbf{w}\|_2^2 = \left\langle \sum_{i=1}^{n} a_i\sigma_i\mathbf{u}_i, \sum_{j=1}^{n} a_j\sigma_j\mathbf{u}_j \right\rangle.$$

Expanding the inner product gives

$$\|A\mathbf{w}\|_2^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} a_ia_j\sigma_i\sigma_j\langle\mathbf{u}_i, \mathbf{u}_j\rangle.$$

Since $\{\mathbf{u}_i\}$ is an orthonormal set, $\langle\mathbf{u}_i, \mathbf{u}_j\rangle = \delta_{ij}$, so all cross terms vanish and we obtain

$$\|A\mathbf{w}\|_2^2 = \sum_{i=1}^{n} a_i^2\sigma_i^2.$$

If $\mathbf{w} \in \text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_{k+1})$, then $a_i = 0$ for $i > k+1$ and $\sum_{i=1}^{k+1} a_i^2 = 1$. Therefore,

$$\|A\mathbf{w}\|_2^2 = \sum_{i=1}^{k+1} a_i^2\sigma_i^2 \geq \sigma_{k+1}^2 \sum_{i=1}^{k+1} a_i^2 = \sigma_{k+1}^2.$$

Since $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \cdots \geq \sigma_{k+1}$, we have $\sigma_i \geq \sigma_{k+1}$ for all $i \leq k+1$.

Finally, taking the square root of both sides, we obtain the lower bound of σ_{k+1} for $\max_{\|\mathbf{x}\|_2=1} \|(A - B_k)\mathbf{x}\|_2$, which is exactly the error obtained from the SVD truncated after k singular values. Therefore, the best possible rank-k approximation for a matrix A for the spectral norm is the SVD of A truncated after k singular values:

$$A_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

3 Data Compression

One key question motivates data compression: given data stored as a matrix, how do we optimize data storage and calculation time while maximizing accuracy? By representing a given matrix through singular value decomposition and applying a low-rank approximation, we achieve exactly that. Modern algorithms further save computational time by computing singular values indirectly, using bidiagonalization and QR factorization to approximate singular values.

3.1 Image Compression

As discussed in the introduction, SVD has many practical applications. In this section, we will explore the practical hurdles involved in calculating SVD through the specific application of SVD to Image Compression. When compressing grayscale images, an $n \times m$ pixel image is stored as $n \times m$ matrix A, where the value of each entry represents the light value of the corresponding pixel. However, by computing the rank k approximation for A, A can be reduced to $U\Sigma V^T$. For the rank k approximation, there are only $n \times k + k + k \times m$ entries (as opposed to the $m \times n$ entries of A).

This reduction is achieved by retaining only the first k columns of U and the first k rows of V^T (the singular vectors corresponding to the largest singular values), while Σ is condensed from a full matrix into a simple list of the k non-zero singular values themselves. With a small enough k, this conversion significantly reduces the amount of data that must be stored. Determining the proper threshold of k, or the number of singular values considered is a practical concern. However, for the purposes of this paper, a simple solution is to arbitrarily dictate a threshold for the minimum singular value.

Another problem arises when computing the SVD of the matrix of an image. Large images are represented as extremely large matrices with the same dimensions of the image itself, and therefore finding the eigenvalues of $A^T A$ is too inefficient. For the direct matrix multiplication required to find $A^T A$, where A is an $m \times n$ matrix, each entry of the n^2 entries in the result require m multiplication operations, and therefore the overall time to compute scales as $O(n^2 m)$. Overall, the time to compute increases roughly cubically with matrix size.

Therefore, numerical libraries for linear algebra opt for algorithms that find the SVD of a matrix without finding its eigenvalues. Many of these algorithms make use of bidiagonalization and QR factorization.

3.2 QR Factorization

QR factorization has been covered in this course based on its connection to the Gram-Schmidt process. That is, any matrix A can be factored as $A = QR$, where Q is an orthogonal and R is upper triangular. The Gram-Schmidt process can be used to obtain a possible QR factorization for A.

3.2.1 The Gram-Schmidt Process

Given a set of linearly independent vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, construct n orthogonal vectors $\mathbf{u}_k$ using the following process:

$$\mathbf{u}_1 = \mathbf{v}_1$$

$$\mathbf{u}_k = \mathbf{v}_k - \sum_{j=1}^{k-1} \text{proj}_{\mathbf{u}_j}(\mathbf{v}_k) = \mathbf{v}_k - \sum_{j=1}^{k-1} \frac{\mathbf{v}_k \cdot \mathbf{u}_j}{||\mathbf{u}_j||^2}\mathbf{u}_j$$

Finally, normalize the set $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ by dividing each vector by their norm to form an orthonormal set $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$, where $\mathbf{e}_k = \frac{\mathbf{u}_k}{||\mathbf{u}_k||}$.

In QR factorization, this orthonormal set obtained from the Gram-Schmidt process forms the columns of the matrix Q. It follows that multiplying Q by R on the right must reverse the Gram-Schmidt process. Hence, the action of R on Q must be:

$$\mathbf{v}_k = ||\mathbf{u}_k||\mathbf{e}_k + \sum_{j=1}^{k-1} \text{proj}_{\mathbf{u}_j}(\mathbf{v}_k) = ||\mathbf{u}_k||\mathbf{e}_k + \sum_{j=1}^{k-1} \frac{\mathbf{v}_k \cdot \mathbf{u}_j}{||\mathbf{u}_j||^2}\mathbf{u}_j$$

We rewrite the summation, using the fact that $\mathbf{e}_k = \frac{\mathbf{u}_k}{||\mathbf{u}_k||}$:

$$\mathbf{v}_k = ||\mathbf{u}_k||\mathbf{e}_k + \sum_{j=1}^{k-1}(\mathbf{v}_k \cdot \mathbf{e}_j)\mathbf{e}_j$$

Finally, we can construct an R such that $A = QR$, and $Q = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n]$

$$A = \begin{bmatrix} | & | & & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \\ | & | & & | \end{bmatrix} = \underbrace{\begin{bmatrix} | & | & & | \\ \mathbf{e}_1 & \mathbf{e}_2 & \cdots & \mathbf{e}_n \\ | & | & & | \end{bmatrix}}_{Q \text{ (orthonormal)}} \underbrace{\begin{bmatrix} \|\mathbf{u}_1\| & \mathbf{e}_1 \cdot \mathbf{v}_2 & \cdots & \mathbf{e}_1 \cdot \mathbf{v}_n \\ 0 & \|\mathbf{u}_2\| & \cdots & \mathbf{e}_2 \cdot \mathbf{v}_n \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \|\mathbf{u}_n\| \end{bmatrix}}_{R \text{ (upper triangular)}}$$

3.2.2 Basic QR Algorithm

The QR algorithm, or QR iteration, is an algorithm for computing the eigenvalues and eigenvectors of a matrix. These definitions are adapted from [4]. Let $A = QR$ be the QR factorization of A, where Q is orthogonal and R is upper triangular. Define a sequence of matrices as follows: Let $A_0 = A$, and $A_k = Q_k R_k$, and $A_{k+1} = R_k Q_k$.

The sequence of matrices are similar, meaning they have the same eigenvalues. Solving for $R_k Q_k$ from $A_k = Q_k R_k$, we get:

$$A_{k+1} = R_kQ_k = Q_k^TA_kQ_k$$

if the QR algorithm converges, then A_k always converges to a matrix with diagonal entries that are the eigenvalues of A.

The basic QR algorithm has a few drawbacks. Firstly, the basic algorithm is inefficient without the application of bidiagonalization, and secondly, with just the basic algorithm, not all matrices have sequences that converge. Modern algorithms increase the efficiency of QR factorization by transforming matrices into *bidiagonal form*. To do so, we enlist the help of another algorithm capable of decomposing a matrix into upper triangular form.

3.3 Householder Transformation

Note: The following definitions in 3.3 are adapted from [5]

Definition 3 (Householder operator)**.** Given an inner product space V with inner product $\langle \cdot, \cdot \rangle$ and a vector $\mathbf{q} \in V$, the *Householder operator* is defined as

$$H_{\mathbf{q}}(\mathbf{x}) = \mathbf{x} - 2\frac{\langle \mathbf{x}, \mathbf{q} \rangle}{\langle \mathbf{q}, \mathbf{q} \rangle}\mathbf{q}.$$

Equivalently, the matrix of a Householder transformation is

$$H_{\mathbf{q}}(\mathbf{x}) = \mathbf{x} - 2\frac{\mathbf{q}^T\mathbf{x}}{\mathbf{q}^T\mathbf{q}}\mathbf{q} = x - 2\frac{\mathbf{q}\mathbf{q}^T}{\mathbf{q}^T\mathbf{q}}\mathbf{x} = (I - 2\frac{\mathbf{q}\mathbf{q}^T}{\mathbf{q}^T\mathbf{q}})x$$

. Note that $(\mathbf{q}^T\mathbf{x})\mathbf{q} = \mathbf{q}(\mathbf{q}^T\mathbf{x}) = \mathbf{q}\mathbf{q}^T\mathbf{x}$ because $\mathbf{q}^T\mathbf{x}$ is a scalar. Also note that the second term is equal to the orthogonal projection of $\mathbf{x}$ onto $\mathbf{q}$.

Geometrically, the Householder transformation reflects a vector $\mathbf{x}$ over a hyperplane (a subspace of dimension one less than its vector space) given a unit vector $\mathbf{q}$ normal to the hyperplane (Fig. 1). A reflection over a hyperplane is equivalent to subtracting $\mathbf{x}$ twice by $\mathbf{x}^{\perp}$, the component of $\mathbf{x}$ perpendicular to the hyperplane. Since $\mathbf{x}^{\perp}$ is equal to the orthogonal projection of $\mathbf{x}$ onto $\mathbf{q}$, the Householder transformation subtracts $\mathbf{x}$ by $2 \cdot \mathrm{proj}_{\mathbf{q}}(\mathbf{x})$.

Corollary 1. *Householder transformations are symmetric and orthogonal in* $\mathbb{R}$*.*

Proof. To show symmetry, we show that $H^T = H$. Applying the transpose to the formula for the matrix,

$$H^T = I^T - (2\frac{\mathbf{q}\mathbf{q}^T}{\mathbf{q}^T\mathbf{q}})^T = I - 2\frac{\mathbf{q}\mathbf{q}^T}{\mathbf{q}^T\mathbf{q}} = H$$

To show that Householder transformations are orthogonal, we show $H^TH = I$:

$$H^TH = H^2 = (I - 2\frac{\mathbf{q}\mathbf{q}^T}{\mathbf{q}^T\mathbf{q}})^2 = I - 4\frac{\mathbf{q}\mathbf{q}^T}{\mathbf{q}^T\mathbf{q}} + 4(\frac{\mathbf{q}\mathbf{q}^T}{\mathbf{q}^T\mathbf{q}})^2$$

Note that $(\frac{\mathbf{q}\mathbf{q}^\mathbf{T}}{\mathbf{q}^\mathbf{T}\mathbf{q}})^2 = \frac{\mathbf{q}\mathbf{q}^\mathbf{T}\mathbf{q}\mathbf{q}^\mathbf{T}}{(\mathbf{q}^\mathbf{T}\mathbf{q})^2} = \frac{\mathbf{q}(\mathbf{q}^\mathbf{T}\mathbf{q})\mathbf{q}^\mathbf{T}}{(\mathbf{q}^\mathbf{T}\mathbf{q})^2} = \frac{\mathbf{q}\mathbf{q}^T}{\mathbf{q}^T\mathbf{q}}$. Returning to the expression for H^TH,

$$I - 4\frac{\mathbf{q}\mathbf{q}^T}{\mathbf{q}^T\mathbf{q}} + 4(\frac{\mathbf{q}\mathbf{q}^T}{\mathbf{q}^T\mathbf{q}})^2 = I$$

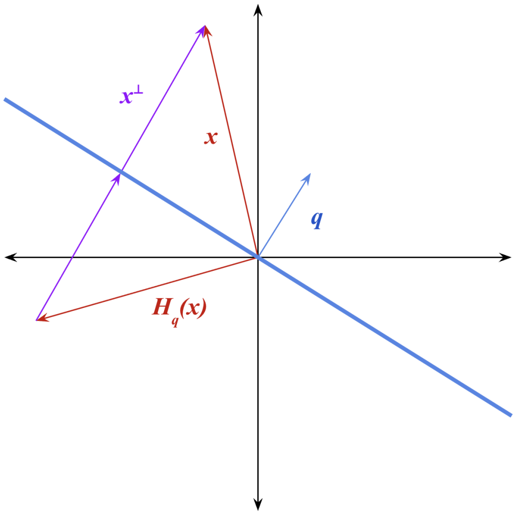

Figure 1: A Householder transformation on a 2-dimensional vector $\mathbf{x}$.

Remark. These results also work for $\mathbb{C}$ if the transpose is replaced with the adjoint. In other words, Householder transformations are unitary.

Therefore, $H^T H = HH^T = I$ and $H^T = H^{-1}$, meaning H is orthogonal.

3.4 QR factorization with Householder Transforms

Recall that an upper triangular matrix is one with nonzero entries only on or above its main diagonal, and that Householder transformations are orthogonal. Now we will show how Householder transformations can be applied to obtain an upper triangular matrix. Let $A \in \mathbb{R}^{m\times n}$ be a real $m \times n$ matrix. We want to inductively show that there exists a sequence of householder operations $H_1, H_2, H_3, \ldots, H_n$ such that $H_n H_{n-1} \ldots H_1 A$ is upper triangular. The idea for this construction follows approach of [6].

For the inductive hypothesis, assume that there exists a sequence of k householder transformations such that

$$H_k H_{k-1} \cdots H_1 A$$

takes the form of a block matrix, with the upper left hand block being an upper triangular matrix U_k:

$$H_k H_{k-1} \cdots H_1 A = \begin{pmatrix} U_k & * \\ 0 & * \end{pmatrix}$$

For our base case, we will let H_1 address the first column of the matrix A, $\mathbf{x}_1$. We want to find a householder transform such that

$$H_1 \mathbf{x}_1 = \beta \mathbf{e}_1$$

where $\mathbf{e}_1$ is the first standard basis vector. Plugging in the definition of the Householder transform, we get:

$$\mathbf{x}_1 - 2\frac{\langle \mathbf{x}_1, \mathbf{q}\rangle}{\langle \mathbf{q}, \mathbf{q}\rangle}\mathbf{q} = \beta \mathbf{e}_1$$

$$2\frac{\langle \mathbf{x}_1, \mathbf{q}\rangle}{\langle \mathbf{q}, \mathbf{q}\rangle}\mathbf{q} = 2\,\mathrm{proj}_{\mathbf{q}}(\mathbf{x}_1) = \mathbf{x}_1 - \beta\mathbf{e}_1$$

$\mathbf{q}$ can be any scalar multiple of $\mathbf{x}_1 - \beta\mathbf{e}_1$. After applying H_1 with $\mathbf{q}$, the first column of A becomes $[\beta, 0, \ldots, 0]^T$:

$$H_1A = \begin{pmatrix} \beta & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & & \vdots \\ 0 & * & \cdots & * \end{pmatrix}$$

This satisfies the form in the inductive hypothesis because the 1×1 matrix β becomes I_1, a 1×1 upper triangular matrix:

$$H_1A = \begin{pmatrix} U_1 & * \\ 0 & * \end{pmatrix}$$

Inductively, we show that if there exists a sequence $H_1 \cdots H_k$ such that we reach the desired block matrix form, then there exists a sequence $H_1 \cdots H_{k+1}$ that satisfies the same hypothesis.

Assume

$$H_k \cdots H_1A = \begin{pmatrix} U_k & * \\ 0 & B \end{pmatrix}$$

where $B \in \mathbb{R}^{(m-k)\times(n-k)}$.

Construct $\hat{H}_{k+1}$ for the $(m-k) \times (n-k)$ submatrix obtained by removing the first k columns and rows of A such that $\hat{H}_{k+1}$ maps the first column of the submatrix to a scalar of the first standard basis vector $\beta_{k+1}e_i$. Construct H_{k+1} as the block matrix

$$H_{k+1} = \begin{pmatrix} I_k & 0 \\ 0 & \hat{H}_{k+1} \end{pmatrix}$$

If we apply H_{k+1} to the block matrix obtained from the sequence $H_1 \cdots H_k$, we get:

$$H_{k+1}H_k \cdots H_1A = \begin{pmatrix} I_k & 0 \\ 0 & \hat{H}_{k+1} \end{pmatrix}\begin{pmatrix} U_k & * \\ 0 & B \end{pmatrix} = \begin{pmatrix} U_k & * \\ 0 & \hat{H}_{k+1}B \end{pmatrix}$$

By construction, $\hat{H}_{k+1}$ maps the first column of B to a multiple of the first standard basis vector, so the upper left block becomes

$$\begin{pmatrix} U_k & 0 \\ 0 & \beta_k + 1 \end{pmatrix} = U_{k+1},$$

and therefore

$$H_{k+1}H_k \cdots H_1A = \begin{pmatrix} I_{k+1} & * \\ 0 & * \end{pmatrix}$$

Therefore, for any finite dimensional A there exists a sequence of Householder Transform matrices such that $H_MH_{M-1}...H_1A$ is upper triangular, where $M = \min(m, n)$. Note that because $H_1, ..., H_M$ are orthogonal, the composition $H_MH_{M-1}...H_1$ is orthogonal. Because Householder transformations are orthogonal and symmetric, each transformation is their own inverse. By multiplying by $H_1H_2...H_M$ on both sides we get A $= H_1H_2...H_MU$. Because U is upper triangular, and the composition of Householder matrices is orthogonal, this is a QR factorization.

3.5 Bidiagonalization

Definition 4 (Bidiagonal matrix)**.** Let A be an $m \times n$ matrix. A is called *bidiagonal* if its only nonzero entries are along the main diagonal and one other diagonal either above or below the main diagonal.

The process of obtaining a bidiagonal factorization is similar to the upper triangular factorization in 5.3.

3.5.1 Golub-Kahan Bidiagonalization

The original algorithm was invented in 1965. The presentation shown follows the outline shown in [7]. Golub-Kahan bidiagonalization makes use of alternating Householder transformations on a matrix to apply the necessary zeroes to each row and column. We will demonstrate the algorithm through induction.

Induction. For an arbitrary integer k, there exist sequences of orthogonal matrices $U_1, \dots, U_k$ and $V_1, \dots, V_k$ such that

$$D = U_k \cdots U_1 \, A \, V_1 \cdots V_k$$

$$D = \begin{pmatrix} B & * \\ 0 & C \end{pmatrix}$$

where B is the upper left $k \times k$ bidiagonal block with entries on the main diagonal and superdiagonal, and C is the $(m-k) \times (n-k)$ trailing submatrix.

Base case. We can take the original matrix as our base case, which represents empty sequences of U and V. Empty sequences of length 0 satisfy the condition trivially, because the empty block matrices satisfy the conditions of bidiagonality and having only zero elements.

For our inductive step, we will construct U_{k+1} and V_{k+1} such that applying U_{k+1} on the left and V_{k+1} on the right yields our inductive hypothesis with a bidiagonal block of $k+1 \times k+1$.

Inductive step. First, define $\hat{U}_{k+1}$ just like in 5.3, so that it maps the first column vector of the $(m-k) \times (n-k)$ submatrix D to $\alpha_{k+1}\mathbf{e}_1$, a scalar multiple of the first standard basis vector of the submatrix. Define U_{k+1} as the block matrix:

$$U_{k+1} = \begin{pmatrix} I_k & 0 \\ 0 & \hat{U}_{k+1} \end{pmatrix}$$

Now, applying U_{k+1} on the left,

$$U_{k+1}D = \begin{pmatrix} I_k & 0 \\ 0 & \hat{U}_{k+1} \end{pmatrix} \begin{pmatrix} B_k & * \\ 0 & C \end{pmatrix} = \begin{pmatrix} B_k & * \\ 0 & \hat{U}_{k+1}C \end{pmatrix}$$

After multiplying by U_{k+1}, the $(k+1)$th row becomes $[0, ..., 0, \alpha_{k+1}, \mathbf{r}^T]$. In order to obtain a bidiagonal upper left block, we want to transform this row to $[0, ..., 0, \alpha_{k+1}, \beta_{k+1}, 0, ..., 0]$.

The desired Householder transformation takes in the $\mathbf{r}^T$ and outputs the row vector $[\beta_{k+1}, 0, ..., 0]$. For finding the Householder transform, we will treat $\mathbf{r}$ as a column vector when multiplying by V_{k+1} on the left. This is possible because $(\mathbf{r}^T\hat{V}_{k+1})^T = \hat{V}_{k+1}^T\mathbf{r}$. Since $\hat{V}$ is symmetric (as a Householder reflector), $\hat{V}^T = \hat{V}$

$$\hat{V}_{k+1}\mathbf{r} = \mathbf{r} - 2\operatorname{proj}_{\mathbf{q}}\mathbf{r} = \alpha_{k+1}\mathbf{e}_1$$

Notice that $\mathbf{q}$ can be any scalar multiple of $r - \alpha_{k+1}\mathbf{e}_1$, so such a householder transform exists. When expressing $\hat{V}_{k+1}$ in matrix form, we embed it in a block matrix as such:

$$V_{k+1} = \begin{pmatrix} I_{k+1} & 0 \\ 0 & \hat{V}_{k+1} \end{pmatrix}.$$

Multiplying $U_{k+1}D$ by V_{k+1} on the right maps the $(k+1)$st row of D starting at the $(k+2)$nd index to a scalar multiple of the first standard basis vector. Hence, after multiplying by U_{k+1} on the left and V_{k+1}, the resulting block matrix can be written as:

$$\begin{pmatrix} B_k & \beta_k & 0 \\ 0 & \alpha_{k+1} & \mathbf{r}^T \\ 0 & 0 & \text{Sub} \end{pmatrix} \begin{pmatrix} I_k & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \hat{V}_{k+1} \end{pmatrix} = \begin{pmatrix} B_k & \beta_k & 0 \\ 0 & \alpha_{k+1} & \mathbf{r}^T V_{k+1} \\ 0 & 0 & \text{Sub}V_{k+1} \end{pmatrix}$$

"Sub" denotes the remaining submatrix, which isn't important in this case.

We know the product $\mathbf{r}^T V_{k+1} = [\beta_{k+1}, 0, ..., 0]$. Hence, after multiplying by U_{k+1} on the left and V_{k+1}, the resulting block matrix can be written as:

$$A_{k+1} = \begin{pmatrix} B_k & \beta_k\mathbf{e}_k & 0 & 0 \\ 0 & \alpha'_{k+1} & \beta'_{k+1} & 0 \\ 0 & 0 & 0 & C' \end{pmatrix}$$

This satisfies the original form of the inductive hypothesis, because it can be written in block matrix form with a $k+1 \times k+1$ bidiagonal matrix as the upper left block, and zeros in the bottom right. Through these constructions, the bidiagonalized upper block can keep expanding until the matrix runs out of rows or columns at which point the final matrix product will be in bidiagonal form. This means the sequences both terminate at $\min(m, n)$.

If $m \geq n$, then the bidiagonalized form will be as follows (All blank space represents zero entries.)

$$\begin{pmatrix} \alpha_1 & \beta_1 & & & & \\ & \alpha_2 & \beta_2 & & & \\ & & \alpha_3 & \beta_3 & & \\ & & & & \ddots & \\ & & & & & \alpha_n \\ 0 & \cdots & & & & \\ \vdots & & & & & \end{pmatrix}$$

and if $m < n$, then the bidiagonalized form will be

$$\begin{pmatrix} \alpha_1 & \beta_1 & & & & \cdots & 0 \\ & \alpha_2 & \beta_2 & & & & \vdots \\ & & \alpha_3 & \beta_3 & & & \\ & & & & \ddots & & \\ & & & & & \alpha_n & \beta_n \end{pmatrix}$$

3.6 Computing SVD using QR Iteration and Bidiagonalization

After bidiagonalizing a matrix, many algorithms/subroutines exist for computing singular values with different data types. According to [8], the standard way to find the singular values of a matrix is to implicitly apply the QR algorithm to $B^T B$, where B is the bidiagonal matrix derived in 3.4. In this way, both bidiagonalization and QR algorithms are heavily utilized in this "standard algorithm." The algorithm we used, which was developed by James Demmel and William Kahan in 1990, is a slight adaptation of this standard algorithm, and therefore also relies on bidiagonalization and QR iteration. However, these algorithms' exact details are beyond the scope of this paper.

While the entire algorithm itself won't be discussed, we will still go over one notable topic used in Demmel and Kahan's method because of its importance and relative simplicity. According to [8], their method makes use of the following rotation matrix:

Definition 5 (Givens rotation)**.** The *Givens rotation* is a rotation matrix with the form

$$G(i, j, \theta) = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & cos(\theta) & \cdots & -sin(\theta) & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & sin(\theta) & \cdots & cos(\theta) & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix}.$$

The nonzero elements are $cos(\theta)$ at $g_{i,i}$, $-sin(\theta)$ at $g_{i,j}$, $sin(\theta)$ at $g_{j,i}$, $cos(\theta)$ at $g_{j,j}$, and 1's for all other diagonal elements.

For any matrix $A_{n\times n}$, $AG(i, j, \theta)$ alters columns i and j, rotating each *subvector* $(g_{k,i}, g_{k,j})$ for $1 \leq k \leq n$ by θ counterclockwise, and $G(i, j, \theta)^T A$ alters rows i and j, rotating each *subvector* $(g_{i,k}, g_{j,k})$ by θ counterclockwise. Furthermore, the action of a Givens matrix on any vector $\mathbf{x}$ in the domain of the linear map A rotates $\mathbf{x}$ counterclockwise by θ [9].

Applying the Givens rotation could annihilate elements in a matrix, setting them to zero:

Example. Consider the matrix

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 6 & 7 \end{bmatrix}.$$

To bring the matrix to upper triangular form, we annihilate $a_{3,2}$. Since the first column already has a pivot, we alter the 2nd and 3rd columns by rotating the subvector $(6, 7)$ to $(0, \sqrt{85})$:

$$AQ(2, 3, \theta) = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 6 & 7 \end{bmatrix} \begin{bmatrix} 1 & * & * \\ 0 & cos(\theta) & -sin(\theta) \\ 0 & sin(\theta) & cos(\theta) \end{bmatrix} = \begin{bmatrix} 1 & * & * \\ 0 & * & * \\ 0 & 0 & \sqrt{85} \end{bmatrix}.$$

Solving for $AQ(2, 3, \theta)$ yields an upper triangular matrix.

Demmel and Kahan devised a way to successively apply Givens rotations to the left and right of the bidiagonal matrix. Iterative "sweeps" of Givens rotations would reduce the values of elements in the upper diagonal, which converge to 0. For a given bidiagonal matrix B, this process results in the decomposition $Q\Sigma P^T$, where Q contains left orthogonal vectors and P contains right orthogonal vectors. Since Q and P are orthogonal, this decomposition preserves the singular values of B as the diagonal in Σ, just as the rotational matrices preserve singular values in SVD.

Using Householder Bidiagonalization and Demmel-Kahan QR iteration, the SciPy SVD method can compute the singular values for any rectangular matrix. This process of first bidiagonalizing using orthogonal Householder transformations, then running QR iteration on the resulting bidiagonal matrix is faster because bidiagonal matrices are sparse, which means most entries of the matrix are zero, so we only need to store and update a few nonzero entries instead of treating it as a full, dense matrix with nonzero entries. Each QR step updates only $O(n)$ nonzeros rather than dense $O(n^2)$ entries, and it avoids explicitly finding A^TA, which is costly for large images.

We implement a simple image compressor, focusing on grayscale images. We set the error threshold $k = 50$: https://github.com/Anwen0507/image-compressor.

4 Conclusion

In this project, we have demonstrated that Singular Value Decomposition is not merely a method for matrix factorization, but a framework for understanding the geometry of linear transformations. By contrasting the algebraic derivation with a topological characterization, we see that singular values serve a dual role: they are simultaneously the square roots of the eigenvalues of A^TA and the measures of maximum geometric distortion induced by the matrix. This geometric insight is what makes the Eckart-Young Theorem intuitive; by retaining the directions of maximum variance (the largest singular values), we guarantee the mathematically optimal approximation of the original data.

While our theoretical exploration focused on the closed form derivation of SVD, the application to image compression reveals the computational constraints of these methods. For large-scale data, the direct computation of eigenvalues for A^TA becomes computationally prohibitive. Consequently, practical implementations rely on iterative algorithms and Golub-Kahan bidiagonalization. Ultimately, SVD bridges the gap between abstract linear algebra and modern data science, providing the rigorous mathematical foundation required for efficient data storage and noise reduction.

References

[1] Robert Friedman. *Honors Math A Notes*. 2010. URL: https://www.math.columbia.edu/~mtwang/teaching/Friedman.pdf.

[2] Lloyd N. Trefethen and David Bau III. *Numerical Linear Algebra*. Philadelphia: SIAM, 1997. ISBN: 978-0-89871-361-9.

[3] Algebraic Pavel (https://math.stackexchange.com/users/90996/algebraic-pavel). *Proof of Eckart-Young-Mirsky theorem*. Mathematics Stack Exchange. URL: https://math.stackexchange.com/questions/759032/proof-of-eckart-young-mirsky-theorem.

[4] Elias Jarlebring. *QR Algorithm*. https://www.math.kth.se/na/SF2524/matber15/qrmethod.pdf. Accessed: 2025-12-21. 2015.

[5] John Kerl. *The Householder Transformation in Numerical Linear Algebra*. Tech. rep. Lecture notes on Householder transforms. Self-published, 2008. URL: https://johnkerl.org/doc/hh.pdf.

[6] Robert van de Geijn and Maggie Myers. *03.3.4 Householder QR factorization, part 1*. YouTube video. Video on Householder QR factorization, posted to YouTube. 2018. URL: https://www.youtube.com/watch?v=5MeeuSoFBdY.

[7] Greg Fasshauer. *12. How to Compute the SVD*. Lecture notes, Numerical Linear Algebra / Computational Mathematics I, Illinois Institute of Technology. 2006. URL: https://www.math.iit.edu/~fass/477577_Chapter_12.pdf.

[8] William Kahan James Demmel. "Accurate Singular Values of Bidiagonal Matrices". In: *SIAM Journal on Scientific Computing* (1990). URL: https://www.netlib.org/lapack/lawnspdf/lawn03.pdf.

[9] "Lecture 13: Givens Rotations". Lecture notes, CS 475/675: Computational Linear Algebra, University of Waterloo. 2025. URL: https://student.cs.uwaterloo.ca/~cs475/CS475-Lecture13.pdf.