and the residual

$$\vec{s}_3 \doteq \vec{a}_3 - \vec{p}_3 = \vec{a}_3 - \vec{q}_1(\vec{q}_1^\top \vec{a}_3) - \vec{q}_2(\vec{q}_2^\top \vec{a}_3). \tag{2.38}$$

These projection formulas only hold because $\{\vec{q}_1, \vec{q}_2\}$ is an orthonormal set. And then we could compute

$$\vec{q}_3 \doteq \frac{\vec{s}_3}{\|\vec{s}_3\|_2}. \tag{2.39}$$

And so on. The general algorithm goes similar.

Algorithm 1 Gram-Schmidt algorithm.

function GRAMSCHMIDTALGORITHM(linearly independent set $\{\vec{a}_1, \ldots, \vec{a}_k\}$)
  $\vec{q}_1 \doteq \vec{a}_1 / \|\vec{a}_1\|_2$.
  for $i \in \{2, 3, \ldots, k\}$ **do**
    $\vec{p}_i \doteq \sum_{j=1}^{i-1} \vec{q}_j(\vec{q}_j^\top \vec{a}_i)$
    $\vec{s}_i \doteq \vec{a}_i - \vec{p}_i$
    $\vec{q}_i \doteq \vec{s}_i / \|\vec{s}_i\|_2$
  end for
  return orthonormal set $\{\vec{q}_1, \ldots, \vec{q}_k\}$
end function

This algorithm has the following two properties, which you can formally prove as an exercise.

Proposition 13 (Gram-Schmidt Algorithm)

Algorithm 1 has the following properties:

1. For each $i \in \{1, \ldots, k\}$, we have

$$\text{span}(\vec{a}_1, \ldots, \vec{a}_i) = \text{span}(\vec{q}_1, \ldots, \vec{q}_i). \tag{2.40}$$

   In particular, $\{\vec{a}_1, \ldots, \vec{a}_k\}$ spans the same subspace as $\{\vec{q}_1, \ldots, \vec{q}_k\}$, as was stated in our original goal.

2. $\{\vec{q}_1, \ldots, \vec{q}_k\}$ is an orthonormal set.

The Gram-Schmidt algorithm leads to something called the QR decomposition. Because, for each i, we have $\text{span}(\vec{a}_1, \ldots, \vec{a}_i) = \text{span}(\vec{q}_1, \ldots, \vec{q}_i)$, it means that we can write $\vec{a}_i$ as a linear combination of the $\vec{q}_j$:

$$\vec{a}_i = r_{1i}\vec{q}_1 + r_{2i}\vec{q}_2 + \cdots + r_{ii}\vec{q}_i = \sum_{j=1}^{i} r_{ji}\vec{q}_j \tag{2.41}$$

Putting all the k equations in a matrix form, we can write

$$\begin{bmatrix} \vec{a}_1 & \cdots & \vec{a}_k \end{bmatrix} = \begin{bmatrix} \vec{q}_1 & \cdots & \vec{q}_k \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1k} \\ 0 & r_{22} & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_{kk} \end{bmatrix}. \tag{2.42}$$

More generally, we can decompose every tall matrix with full column rank into a product of a tall matrix with orthonormal columns Q and an upper-triangular matrix R.

Theorem 14 (QR Decomposition)
Let $A \in \mathbb{R}^{n \times k}$ where $k \leq n$ (so A is tall). Suppose A has full column rank. Then there is a matrix $Q \in \mathbb{R}^{n \times k}$ with orthonormal columns, and a matrix $R \in \mathbb{R}^{k \times k}$ which is upper triangular, such that $A = QR$.

As a final note, there are various alterations to the QR decomposition that work for matrices which are wide and/or do not have full column rank. Those are out of scope, but the idea is the same.

The QR decomposition is also relevant in numerical linear algebra, where it can be used to solve tall linear systems $A\vec{x} = \vec{y}$ efficiently, especially if the underlying matrix A has special structure. All such connections are out of scope.

2.3 Fundamental Theorem of Linear Algebra

The fundamental theorem of linear algebra is a tool for understanding what happens to vectors and vector spaces under a linear transformation. Matrix multiplication transforms one vector space into another. This is helpful for allowing us to change our coordinate system, which tells us more about the problem.

Definition 15 (Direct Sum)
Let $U, V \subseteq \mathbb{R}^n$ be subspaces. We say that U and V *direct sum* to $\mathbb{R}^n$, denoted $U \oplus V = \mathbb{R}^n$, if and only if:

- Every vector $\vec{x} \in \mathbb{R}^n$ can be written as $\vec{x} = \vec{x}_1 + \vec{x}_2$, where $\vec{x}_1 \in U$ and $\vec{x}_2 \in V$.
- Furthermore, this decomposition is unique, in the sense that if $\vec{x} = \vec{x}_1 + \vec{x}_2 = \vec{y}_1 + \vec{y}_2$ are two instances of the above decomposition, then $\vec{x}_1 = \vec{y}_1$ and $\vec{x}_2 = \vec{y}_2$.

Theorem 16 (Fundamental Theorem of Linear Algebra)
Let $A \in \mathbb{R}^{m \times n}$. Then

$$\mathcal{N}(A) \oplus \mathcal{R}\left(A^\top\right) = \mathbb{R}^n. \tag{2.43}$$

Note that we *cannot* replace $\mathcal{R}\left(A^\top\right)$ by $\mathcal{R}(A)$, since vectors in $\mathcal{R}(A)$ and $\mathcal{N}(A)$ do not even have the same number of entries or lie in the same Euclidean space. If we want to make a statement about $\mathcal{R}(A)$, we can replace A by $A^\top$ in the above theorem to get the following corollary.

Corollary 17. *Let $A \in \mathbb{R}^{m \times n}$. Then*

$$\mathcal{N}\left(A^\top\right) \oplus \mathcal{R}(A) = \mathbb{R}^m. \tag{2.44}$$

To prove the fundamental theorem of linear algebra, we use a tool called the *orthogonal decomposition theorem*.

Definition 18 (Orthogonal Complement)
Let $S \subseteq \mathbb{R}^n$ be a subspace. The *orthogonal complement* of S, denoted $S^\perp$, is

$$S^\perp \doteq \{\vec{x} \in \mathbb{R}^n \mid \vec{s}^\top \vec{x} = 0 \text{ for all } \vec{s} \in S\} \tag{2.45}$$

Theorem 19 (Orthogonal Decomposition Theorem, Theorem 2.1 of [2])
Let $S \subseteq \mathbb{R}^n$ be a subspace. Then

$$S \oplus S^\perp = \mathbb{R}^n. \tag{2.46}$$

Proof. To prove this, we first need to prove the following claim:

> Let $U, V \subseteq \mathbb{R}^n$ be subspaces. Then $U \oplus V = \mathbb{R}^n$ if and only if every vector $\vec{x} \in \mathbb{R}^n$ can be written as $\vec{x} = \vec{x}_1 + \vec{x}_2$, where $\vec{x}_1 \in U$ and $\vec{x}_2 \in V$, and $U \cap V = \{\vec{0}\}$.

To prove this claim, suppose first that $U \oplus V = \mathbb{R}^n$. Then every every vector $\vec{x} \in \mathbb{R}^n$ can be written as $\vec{x} = \vec{x}_1 + \vec{x}_2$, where $\vec{x}_1 \in U$ and $\vec{x}_2 \in V$. It remains to prove that $U \cap V = \{\vec{0}\}$. Suppose for the sake of contradiction that there exists $\vec{y} \neq \vec{0}$ such that $\vec{y} \in U \cap V$. Then

$$\vec{x} = (\vec{x}_1 + \vec{y}) + (\vec{x}_2 - \vec{y}). \tag{2.47}$$

Since $\vec{y} \in U$, we have $\vec{x}_1 + \vec{y} \in U$; since $\vec{y} \in V$, we have $\vec{x}_2 - \vec{y} \in V$. Thus

$$\vec{x} = \vec{x}_1 + \vec{x}_2 = (\vec{x}_1 + \vec{y}) + (\vec{x}_2 - \vec{y}) \tag{2.48}$$

are two distinct ways to write $\vec{x}$ as the sum of vectors from U and V, so it cannot be true that $U \oplus V = \mathbb{R}^n$, a contradiction.

Towards the other direction, suppose that every vector $\vec{x} \in \mathbb{R}^n$ can be written as $\vec{x} = \vec{x}_1 + \vec{x}_2$, where $\vec{x}_1 \in U$ and $\vec{x}_2 \in V$, and $U \cap V = \{\vec{0}\}$. The only thing remaining to prove is that if

$$\vec{x} = \vec{x}_1 + \vec{x}_2 = \vec{z}_1 + \vec{z}_2 \tag{2.49}$$

where $\vec{x}_1, \vec{z}_1 \in U$ and $\vec{x}_2, \vec{z}_2 \in V$, then we must have $\vec{x}_1 = \vec{z}_1$ and $\vec{x}_2 = \vec{z}_2$. Suppose again for the sake of contradiction that there exists $\vec{x} \in \mathbb{R}^n$, $\vec{x}_1, \vec{z}_1 \in U$, and $\vec{x}_2, \vec{z}_2 \in V$ such that

$$\vec{x} = \vec{x}_1 + \vec{x}_2 = \vec{z}_1 + \vec{z}_2 \tag{2.50}$$

but $\vec{x}_1 \neq \vec{z}_1$ or $\vec{x}_2 \neq \vec{z}_2$. Then we have

$$\vec{0} = \vec{x} - \vec{x} = \vec{x}_1 + \vec{x}_2 - \vec{z}_1 - \vec{z}_2 = (\vec{x}_1 - \vec{z}_1) + (\vec{x}_2 - \vec{z}_2). \tag{2.51}$$

Thus, we have that

$$\vec{x}_1 - \vec{z}_1 = \vec{z}_2 - \vec{x}_2 \neq \vec{0}. \tag{2.52}$$

Since $\vec{x}_1, \vec{z}_1 \in U$, we have $\vec{x}_1 - \vec{z}_1 \in U$, and since $\vec{x}_2, \vec{z}_2 \in V$, we have $\vec{z}_2 - \vec{x}_2 \in V$. Since they are equal, we have $\vec{x}_1 - \vec{z}_1 \in U \cap V$ and nonzero. Thus $U \cap V \neq \{\vec{0}\}$, a contradiction.

This proves the above claim. Now to prove the actual theorem, we note that every vector $\vec{x} \in \mathbb{R}^n$ can be written as

$$\vec{x} = \text{proj}_S(\vec{x}) + (\vec{x} - \text{proj}_S(\vec{x})). \tag{2.53}$$

By definition, $\text{proj}_S(\vec{x}) \in S$, and because the projection residual is orthogonal to the subspace, we have $\vec{x} - \text{proj}_S(\vec{x}) \in S^\perp$. Thus every vector in $\mathbb{R}^n$ can be written as the sum of a vector in S and $S^\perp$. It is an exercise to show that $S \cap S^\perp = \{\vec{0}\}$. Invoking the quoted claim completes the proof. □

Using this theorem, the only thing we need to show to prove the fundamental theorem of linear algebra is that $\mathcal{N}(A)$ and $\mathcal{R}(A^\top)$ are orthogonal complements. We do this below.

Proof of Theorem 16. By Theorem 19, the only thing we need to show is that $\mathcal{N}(A) = \mathcal{R}\left(A^\top\right)^\perp$. This is a set equality; we show it by showing that $\mathcal{N}(A) \subseteq \mathcal{R}\left(A^\top\right)^\perp$ and that $\mathcal{N}(A) \supseteq \mathcal{R}\left(A^\top\right)^\perp$.

We first want to show that $\mathcal{N}(A) \subseteq \mathcal{R}\left(A^\top\right)^\perp$. That is, we want to show that for any $\vec{x} \in \mathcal{N}(A)$ we have $\vec{x} \in \mathcal{R}\left(A^\top\right)^\perp$. That is, for any $\vec{y} \in \mathcal{R}\left(A^\top\right)$, we want to show that $\vec{y}^\top \vec{x} = 0$.

Since $\vec{y} \in \mathcal{R}\left(A^\top\right)$ we can write $\vec{y} = A^\top \vec{w}$ for some $\vec{w} \in \mathbb{R}^m$. Then, since $\vec{x} \in \mathcal{N}(A)$ we have $A\vec{x} = \vec{0}$, so

$$\begin{aligned} \vec{y}^\top \vec{x} &= (A^\top \vec{w})^\top \vec{x} && (2.54)\\ &= \vec{w}^\top A\vec{x} && (2.55)\\ &= \vec{w}^\top \vec{0} && (2.56)\\ &= 0. && (2.57) \end{aligned}$$

Thus $\vec{x}$ and $\vec{y}$ are orthogonal, so $\vec{x} \in \mathcal{R}\left(A^\top\right)^\perp$, which shows that $\mathcal{N}(A) \subseteq \mathcal{R}\left(A^\top\right)^\perp$.

We now want to show that $\mathcal{R}\left(A^\top\right)^\perp \subseteq \mathcal{N}(A)$. That is, we want to show that for any $\vec{x} \in \mathcal{R}\left(A^\top\right)^\perp$, we want to show that $\vec{x} \in \mathcal{N}(A)$. That is, we want to show that $A\vec{x} = \vec{0}$.

By definition, for every $\vec{y} \in \mathcal{R}\left(A^\top\right)$, we have $\vec{y}^\top \vec{x} = 0$. By writing $\vec{y} = A^\top \vec{w}$ for arbitrary $\vec{w} \in \mathbb{R}^m$, we get that for every $\vec{w} \in \mathbb{R}^m$ we have $(A^\top \vec{w})^\top \vec{x} = 0$. But the left-hand side is $\vec{w}^\top A\vec{x}$, so we have that $\vec{w}^\top A\vec{x} = 0$ for every $\vec{w} \in \mathbb{R}^m$. This is true for all $\vec{w} \in \mathbb{R}^m$, so it is true for the specific choice of $\vec{w} = A\vec{x}$, which yields

$$\begin{aligned} 0 &= \vec{w}^\top A\vec{x} && (2.58)\\ &= (A\vec{x})^\top A\vec{x} && (2.59)\\ &= \|A\vec{x}\|_2^2 && (2.60)\\ \implies A\vec{x} &= \vec{0}. && (2.61) \end{aligned}$$

This implies that $\vec{x} \in \mathcal{N}(A)$ as desired, so $\mathcal{R}\left(A^\top\right)^\perp \subseteq \mathcal{N}(A)$.

Thus, we have shown that $\mathcal{N}(A) = \mathcal{R}\left(A^\top\right)^\perp$, and so by Theorem 19 we have $\mathcal{N}(A) \oplus \mathcal{R}\left(A^\top\right) = \mathbb{R}^n$. □

This will help us solve a very important optimization problem, which is considered "dual" to least squares in some sense. Recall that least squares helps us find an approximate solution to the linear system $A\vec{x} = \vec{y}$, when A is a *tall* matrix with full column rank. In other words, the linear system is over-determined, there are many more equations than unknowns, and there are generally no exact solutions, so we pick the solution with minimum squared error.

What about when A is a *wide* matrix with full *row* rank? There are now more unknowns than equations, and infinitely many exact solutions. So how do we pick one solution in particular? It really depends on which engineering problem we are solving. One *common* solution is to pick the minimum-energy or minimum-norm problem, which is the solution to the optimization problem:

$$\begin{aligned} \min_{\vec{x} \in \mathbb{R}^n} \quad & \|\vec{x}\|_2^2 \\ \text{s.t.} \quad & A\vec{x} = \vec{y}. \end{aligned} \tag{2.62}$$

Note that this principle of choosing the smallest or simplest solution — the "Occam's Razor" principle — is much more broadly generalized beyond the case of finding solutions to linear systems, and is used within control theory and machine learning. But we deal with just this linear system case for now.

Theorem 20 (Minimum-Norm Solution)

Let $A \in \mathbb{R}^{m \times n}$ have full row rank, and let $\vec{y} \in \mathbb{R}^m$. Then the solution to Equation (2.62), i.e., the solution to

$$\begin{aligned} \min_{\vec{x} \in \mathbb{R}^n} \quad & \|\vec{x}\|_2^2 \\ \text{s.t.} \quad & A\vec{x} = \vec{y}, \end{aligned} \tag{2.62}$$

is given by

$$\vec{x}^\star = A^\top (AA^\top)^{-1} \vec{y}. \tag{2.63}$$

Proof. Observe that the constraint $A\vec{x} = \vec{y}$ under-specifies the $\vec{x}$ — in particular, any component of $\vec{x}$ in $\mathcal{N}(A)$ will not affect the constraint and only the objective. In this sense, it is "wasteful", and we should intuitively remove it. This motivates using Theorem 16 to decompose $\vec{x}$ into a component inside $\mathcal{N}(A)$ — which we want to remove — and a component inside $\mathcal{R}\left(A^\top\right)$ — which we will optimize over.

Indeed, write $\vec{x} = \vec{u} + \vec{v}$, where $\vec{u} \in \mathcal{N}(A)$ and $\vec{v} \in \mathcal{R}\left(A^\top\right)$. Thus, there exists $\vec{w} \in \mathbb{R}^m$ such that $\vec{v} = A^\top \vec{w}$. The constraint becomes

$$\begin{aligned} \vec{y} &= A\vec{x} && (2.64)\\ &= A(\vec{u} + \vec{v}) && (2.65)\\ &= A\vec{u} + A\vec{v} && (2.66)\\ &= \vec{0} + AA^\top \vec{w} && (2.67)\\ &= AA^\top \vec{w}. && (2.68) \end{aligned}$$

And the objective function becomes

$$\begin{aligned} \|\vec{x}\|_2^2 &= \|\vec{u} + \vec{v}\|_2^2 && (2.69)\\ &= \vec{u}^\top \vec{u} + 2\vec{u}^\top \vec{v} + \vec{v}^\top \vec{v} && (2.70)\\ &= \|\vec{u}\|_2^2 + 2\vec{v}^\top \vec{u} + \|\vec{v}\|_2^2 && (2.71)\\ &= \|\vec{u}\|_2^2 + 2(A^\top \vec{w})^\top \vec{u} + \|\vec{v}\|_2^2 && (2.72)\\ &= \|\vec{u}\|_2^2 + 2\vec{w}^\top A\vec{u} + \|\vec{v}\|_2^2 && (2.73)\\ &= \|\vec{u}\|_2^2 + 2\vec{w}^\top \vec{0} + \|\vec{v}\|_2^2 && (2.74)\\ &= \|\vec{u}\|_2^2 + 2 \cdot 0 + \|\vec{v}\|_2^2 && (2.75)\\ &= \|\vec{u}\|_2^2 + \|\vec{v}\|_2^2 && (2.76)\\ &= \|\vec{u}\|_2^2 + \left\|A^\top \vec{w}\right\|_2^2 && (2.77) \end{aligned}$$

Thus, the minimum-norm problem can be reformulated in terms of $\vec{u}$ and $\vec{w}$:

$$\begin{aligned} \min_{\substack{\vec{u} \in \mathbb{R}^n \\ \vec{w} \in \mathbb{R}^m}} \quad & \|\vec{u}\|_2^2 + \left\|A^\top \vec{w}\right\|_2^2 \\ \text{s.t.} \quad & \vec{y} = AA^\top \vec{w} \\ & A\vec{u} = \vec{0}. \end{aligned} \tag{2.78}$$

Now, because A has full row rank, $AA^\top$ is invertible, so the first constraint implies that $\vec{w}^\star = (AA^\top)^{-1}\vec{y}$, so $\vec{v}^\star = A^\top \vec{w}^\star = A^\top (AA^\top)^{-1}\vec{y}$. And because we are trying to minimize the objective, which only involves $\vec{u}$ through $\|\vec{u}\|_2^2$, the ideal solution is to set $\vec{u}^\star = \vec{0}$, which also satisfies the second constraint and so is feasible. Thus $\vec{x}^\star = \vec{v}^\star = A^\top (AA^\top)^{-1}\vec{y}$ as desired. □

2.4 Symmetric Matrices

Symmetric matrices are a sub-class of matrices which have many special properties, and in engineering applications one usually tries to work with symmetric matrices as much as possible.

Definition 21 (Symmetric Matrix)
Let $A \in \mathbb{R}^{n\times n}$ be a square matrix. We say that A is *symmetric* if $A = A^\top$. The set of all symmetric matrices is denoted $\mathbb{S}^n$.

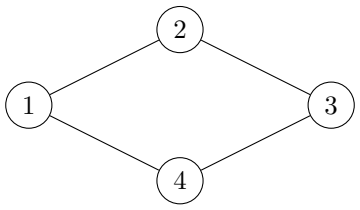
Equivalently, $A_{ij} = A_{ji}$ for all i and j.

Example 22. The 2×2 matrix $\begin{bmatrix} a & b \\ b & c \end{bmatrix}$ is symmetric.

Example 23 (Covariance Matrices)**.** Any matrix of the form $A = BB^\top$, such as the covariance matrices we will discuss in the next section, is a symmetric matrix, since

$$A^\top = (BB^\top)^\top = (B^\top)^\top (B)^\top = BB^\top = A. \tag{2.79}$$

Example 24 (Adjacency Matrix)**.** Consider an *undirected* connected graph $G = (V, E)$, for example the following:



Its adjacency matrix A has coordinate $A_{ij} = 1$ if $(i, j) \in E$, and $A_{ij} = 0$ otherwise; in the above example, we have

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}. \tag{2.80}$$

Since the graph is undirected, $(i, j) \in E$ if and only if $(j, i) \in E$, so $A_{ij} = A_{ji}$, and so A is a symmetric matrix.

Why do we care about symmetric matrices? Symmetric matrices have two nice properties: real eigenvalues, and guaranteed diagonalizability.

In general, a (non-symmetric) matrix need not be diagonalizable. For example, the matrix $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ is not diagonalizable. How can we characterize the diagonalizability of a matrix, then?

First, we will need the following definitions.

Definition 25 (Multiplicities)
Let $A \in \mathbb{R}^{n\times n}$, and let λ be an eigenvalue of A.

(a) The *algebraic multiplicity* μ of eigenvalue λ in A is the number of times λ is a root of the characteristic polynomial $p_A(x) \doteq \det(xI - A)$ of A, i.e., it is the power of $(x - \lambda)$ in the factorization of $p_A(x)$.

(b) The *geometric multiplicity* ϕ of eigenvalue λ in A is the dimension of the null space $\Phi \doteq \mathcal{N}(\lambda I - A)$.

Theorem 26 (Diagonalizability)
A square matrix $A \in \mathbb{R}^{n \times n}$ is diagonalizable if and only if every eigenvalue of A has equal algebraic and geometric multiplicities.

Example 27 (Multiplicities of Degenerate Matrix)**.** We were earlier told that the matrix $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ is not diagonalizable. To check this, let us compute its eigenvalues, algebraic multiplicities, and geometric multiplicities.

First, its characteristic polynomial is

$$p_A(x) = \det(xI - A) \tag{2.81}$$
$$= \det\left(\begin{bmatrix} x-1 & -1 \\ 0 & x-1 \end{bmatrix}\right) \tag{2.82}$$
$$= (x-1)^2. \tag{2.83}$$

Thus, A has only one eigenvalue $\lambda = 1$. Since $(x-1)$ has power 2 in the factorization of p_A, the eigenvalue $\lambda = 1$ has algebraic multiplicity $\mu = 2$.

The corresponding null space is

$$\Phi = \mathcal{N}(\lambda I - A) \tag{2.84}$$
$$= \mathcal{N}\left(\begin{bmatrix} 1-1 & -1 \\ 0 & 1-1 \end{bmatrix}\right) \tag{2.85}$$
$$= \mathcal{N}\left(\begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix}\right) \tag{2.86}$$
$$= \operatorname{span}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) \tag{2.87}$$

which has dimension $\phi = 1$. Thus, for $\lambda = 1$, we have $\mu \neq \phi$ and the matrix is indeed not diagonalizable.

This allows us to formally state the spectral theorem.

Theorem 28 (Spectral Theorem)
Let $A \in \mathbb{S}^n$ have eigenvalues λ_i with algebraic multiplicities μ_i, eigenspaces $\Phi_i \doteq \mathcal{N}(\lambda_i I - A)$, and geometric multiplicities $\phi_i \doteq \dim(\Phi_i)$.

(a) All eigenvalues are real: $\lambda_i \in \mathbb{R}$ for each i.

(b) Eigenspaces corresponding to different eigenvalues are orthogonal: Φ_i and Φ_j are orthogonal subspaces, i.e., for every $\vec{p}_i \in \Phi_i$ and $\vec{p}_j \in \Phi_j$ we have $\vec{p}_i^{\top} \vec{p}_j = 0$.

(c) A is diagonalizable: $\mu_i = \phi_i$ for each i.

(d) A is *orthonormally diagonalizable*; there exists an *orthonormal* matrix $U \in \mathbb{R}^{n \times n}$ and *diagonal* matrix $\Lambda \in \mathbb{R}^{n \times n}$ such that $A = U \Lambda U^{\top}$.

Recall that orthonormal matrices are matrices whose columns are orthonormal, i.e., are pairwise orthogonal and unit-norm. Orthonormal matrices U have the nice property that $U^\top U = I$, and if U is square, then $U^\top = U^{-1}$.

Proof of Theorem 28. Part (a) might be left to homework; part (b) will definitely be left to homework; we prove parts (c) and (d) here. In particular, we assume that parts (a) and (b) are true, and attempt to prove (d). Note that (d) implies (c), as an orthonormal diagonalization is a type of diagonalization, and so the existence of an orthonormal diagonalization must require the algebraic and geometric multiplicities to be equal.

Our proof strategy is to use induction on n, the size of the matrix. The base case of our induction is 1×1 matrices, for which the diagonalization is trivial. Now consider the inductive step. Our hope is, given $A \in \mathbb{S}^n$ which has eigenvalue λ, to get a decomposition of the form

$$A = V \begin{bmatrix} \lambda & \vec{0}^\top \\ \vec{0} & B \end{bmatrix} V^\top \qquad \text{or equivalently} \qquad V^\top A V = \begin{bmatrix} \lambda & \vec{0}^\top \\ \vec{0} & B \end{bmatrix} \tag{2.88}$$

where $V \in \mathbb{R}^{n \times n}$ is orthonormal and $B \in \mathbb{S}^{n-1}$ is symmetric. If we can do that, then we can inductively diagonalize $B = W \Gamma W^\top$, and finally use that to construct a diagonalization for $A = U \Lambda U^\top$, where $U \in \mathbb{R}^{n \times n}$ is orthonormal and $\Lambda \doteq \begin{bmatrix} \lambda & \vec{0}^\top \\ \vec{0} & \Gamma \end{bmatrix}$.

Let $\vec{u}$ be a unit-norm eigenvector of A corresponding to eigenvalue λ. Remember that we want an orthonormal matrix $V \in \mathbb{R}^{n \times n}$ which "isolates" λ. This motivates using $\vec{u}$ and a basis of the orthogonal complement of $\text{span}(\vec{u})$ to form V. To construct this matrix V, we run Gram-Schmidt on the columns of the matrix $\begin{bmatrix} \vec{u} & I \end{bmatrix} \in \mathbb{R}^{n \times (n+1)}$, throwing out the single vector which will have 0 projection residual (there must be exactly one such vector by a counting argument; to get n linearly independent vectors from a spanning set of $n+1$ vectors, we need to remove exactly one vector), and obtaining the orthonormal matrix $V = \begin{bmatrix} \vec{u} & V_1 \end{bmatrix} \in \mathbb{R}^{n \times n}$ where $V_1 \in \mathbb{R}^{n \times (n-1)}$ is itself orthonormal. By construction, we have $V_1^\top \vec{u} = \vec{0}$ and $\vec{u}^\top V_1 = \vec{0}^\top$. Thus,

$$\begin{aligned}
V^\top A V &= \begin{bmatrix} \vec{u} & V_1 \end{bmatrix}^\top A \begin{bmatrix} \vec{u} & V_1 \end{bmatrix} && (2.89)\\
&= \begin{bmatrix} \vec{u}^\top \\ V_1^\top \end{bmatrix} A \begin{bmatrix} \vec{u} & V_1 \end{bmatrix} && (2.90)\\
&= \begin{bmatrix} \vec{u}^\top \\ V_1^\top \end{bmatrix} \begin{bmatrix} A\vec{u} & AV_1 \end{bmatrix} && (2.91)\\
&= \begin{bmatrix} \vec{u}^\top \\ V_1^\top \end{bmatrix} \begin{bmatrix} \lambda\vec{u} & AV_1 \end{bmatrix} && (2.92)\\
&= \begin{bmatrix} \lambda\vec{u}^\top\vec{u} & \vec{u}^\top AV_1 \\ \lambda V_1^\top\vec{u} & V_1^\top AV_1 \end{bmatrix} && (2.93)\\
&= \begin{bmatrix} \lambda \|\vec{u}\|_2^2 & (A^\top\vec{u})^\top V_1 \\ \vec{0} & V_1^\top AV_1 \end{bmatrix} && (2.94)\\
&= \begin{bmatrix} \lambda & (A\vec{u})^\top V_1 \\ \vec{0} & V_1^\top AV_1 \end{bmatrix} && (2.95)\\
&= \begin{bmatrix} \lambda & \lambda\vec{u}^\top V_1 \\ \vec{0} & V_1^\top AV_1 \end{bmatrix} && (2.96)\\
&= \begin{bmatrix} \lambda & \vec{0}^\top \\ \vec{0} & V_1^\top AV_1 \end{bmatrix} && (2.97)
\end{aligned}$$

$$= \begin{bmatrix} \lambda & \vec{0}^\top \\ \vec{0} & B \end{bmatrix}, \tag{2.98}$$

where $B \doteq V_1^\top A V_1$ in accordance with our proof outline. Now we need to check that B is symmetric; indeed, we have

$$B^\top = (V_1^\top A V_1)^\top \tag{2.99}$$
$$= (V_1)^\top A^\top (V_1^\top)^\top \tag{2.100}$$
$$= V_1^\top A V_1 \tag{2.101}$$
$$= B. \tag{2.102}$$

By induction, we can orthonormally diagonalize this matrix as $B = W \Gamma W^\top \in \mathbb{R}^{(n-1)\times(n-1)}$, where $W \in \mathbb{R}^{(n-1)\times(n-1)}$ is orthonormal and $\Gamma \in \mathbb{R}^{(n-1)\times(n-1)}$ is diagonal. Thus, by using $W^{-1} = W^\top$, we have

$$\Gamma = W^\top B W \tag{2.103}$$
$$= W^\top V_1^\top A V_1 W \tag{2.104}$$
$$= (V_1 W)^\top A (V_1 W). \tag{2.105}$$

We want an orthonormal matrix $U \in \mathbb{R}^{n\times n}$ such that $U^\top A U = \Lambda = \begin{bmatrix} \lambda & \vec{0}^\top \\ \vec{0} & \Gamma \end{bmatrix}$. Thus, the above calculation motivates the choice $U = \begin{bmatrix} \vec{u} & V_1 W \end{bmatrix} \in \mathbb{R}^{n\times n}$. Thus

$$U^\top A U = \begin{bmatrix} \vec{u} & V_1 W \end{bmatrix}^\top A \begin{bmatrix} \vec{u} & V_1 W \end{bmatrix} \tag{2.106}$$
$$= \begin{bmatrix} \vec{u}^\top \\ W^\top V_1^\top \end{bmatrix} A \begin{bmatrix} \vec{u} & V_1 W \end{bmatrix} \tag{2.107}$$
$$= \begin{bmatrix} \vec{u}^\top \\ W^\top V_1^\top \end{bmatrix} \begin{bmatrix} A\vec{u} & A V_1 W \end{bmatrix} \tag{2.108}$$
$$= \begin{bmatrix} \vec{u}^\top \\ W^\top V_1^\top \end{bmatrix} \begin{bmatrix} \lambda\vec{u} & A V_1 W \end{bmatrix} \tag{2.109}$$
$$= \begin{bmatrix} \lambda \vec{u}^\top \vec{u} & \vec{u}^\top A V_1 W \\ \lambda W^\top V_1^\top \vec{u} & W^\top V_1^\top A V_1 W \end{bmatrix} \tag{2.110}$$
$$= \begin{bmatrix} \lambda \|\vec{u}\|_2^2 & (A^\top \vec{u})^\top V_1 W \\ \lambda W^\top \vec{0} & W^\top B W \end{bmatrix} \tag{2.111}$$
$$= \begin{bmatrix} \lambda & (A\vec{u})^\top V_1 W \\ \vec{0} & \Gamma \end{bmatrix} \tag{2.112}$$
$$= \begin{bmatrix} \lambda & \lambda \vec{u}^\top V_1 W \\ \vec{0} & \Gamma \end{bmatrix} \tag{2.113}$$
$$= \begin{bmatrix} \lambda & \lambda \vec{0}^\top W \\ \vec{0} & \Gamma \end{bmatrix} \tag{2.114}$$
$$= \begin{bmatrix} \lambda & \vec{0}^\top \\ \vec{0} & \Gamma \end{bmatrix} \tag{2.115}$$
$$= \Lambda, \tag{2.116}$$

as desired. Thus $A = U \Lambda U^\top$ is an orthonormal diagonalization of A. This proves (d), and hence (c). □

One nice thing about diagonalization is that we can read off the eigenvalues and eigenvectors from the components of the diagonalization.

Proposition 29

Let $A \in \mathbb{S}^n$ have orthonormal diagonalization $A = U\Lambda U^\top$, where $U = \begin{bmatrix} \vec{u}_1 & \cdots & \vec{u}_n \end{bmatrix} \in \mathbb{R}^{n\times n}$ is square orthonormal, and $\Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \in \mathbb{R}^{n\times n}$ is diagonal. Then for each i, the pair $(\lambda_i, \vec{u}_i)$ is an eigenvalue-eigenvector pair for A.

Proof. By using $U^\top = U^{-1}$, we have

$$A = U\Lambda U^\top \tag{2.117}$$

$$AU = U\Lambda \tag{2.118}$$

$$A\begin{bmatrix} \vec{u}_1 & \cdots & \vec{u}_n \end{bmatrix} = \begin{bmatrix} \vec{u}_1 & \cdots & \vec{u}_n \end{bmatrix}\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \tag{2.119}$$

$$\begin{bmatrix} A\vec{u}_1 & \cdots & A\vec{u}_n \end{bmatrix} = \begin{bmatrix} \lambda_1\vec{u}_1 & \cdots & \lambda_n\vec{u}_n \end{bmatrix}. \tag{2.120}$$

Therefore, each $(\lambda_i, \vec{u}_i)$ is an eigenvalue-eigenvector pair of A. □

Using this, we can work with another nice property of the orthonormal diagonalization. Namely, we can read off bases for $\mathcal{N}(A)$ and $\mathcal{R}(A)$. That is, a basis for $\mathcal{N}(A)$ is the set of eigenvectors $\vec{u}_i$ corresponding to the eigenvalues λ_i of A which are equal to 0. Since U is orthonormal, the remaining eigenvectors $\vec{u}_i$ span the orthogonal complement to $\mathcal{N}(A)$. But by the fundamental theorem of linear algebra (Theorem 16), we have $\mathcal{N}(A)^\perp = \mathcal{R}(A^\top) = \mathcal{R}(A)$, so these eigenvectors form a basis for $\mathcal{R}(A)$. Soon, we'll discover the singular value decomposition, which allows for this kind of decomposition of a matrix into its range and null spaces, except for arbitrary matrices.

Before we get into those, we will first state and solve a quick optimization problem which yields the eigenvalues of a symmetric matrix. This optimization problem turns out to be quite useful for further study of optimization.

Theorem 30 (Variational Characterization of Eigenvalues)

Let $A \in \mathbb{S}^n$. Let $\lambda_{\min}\{A\}$ and $\lambda_{\max}\{A\}$ be the maximum and minimum eigenvalues of A (which is well-defined since by the spectral theorem, all eigenvalues of A are real). Then

$$\lambda_{\max}\{A\} = \max_{\substack{\vec{x}\in\mathbb{R}^n \\ \vec{x}\neq\vec{0}}} \frac{\vec{x}^\top A\vec{x}}{\vec{x}^\top\vec{x}} = \max_{\substack{\vec{x}\in\mathbb{R}^n \\ \|\vec{x}\|_2=1}} \vec{x}^\top A\vec{x} \tag{2.121}$$

$$\lambda_{\min}\{A\} = \min_{\substack{\vec{x}\in\mathbb{R}^n \\ \vec{x}\neq\vec{0}}} \frac{\vec{x}^\top A\vec{x}}{\vec{x}^\top\vec{x}} = \min_{\substack{\vec{x}\in\mathbb{R}^n \\ \|\vec{x}\|_2=1}} \vec{x}^\top A\vec{x}. \tag{2.122}$$

The term $\dfrac{\vec{x}^\top A\vec{x}}{\vec{x}^\top\vec{x}}$ is called the *Rayleigh quotient* of A; it is a function of $\vec{x} \in \mathbb{R}^n$.

Proof. Before we start trying to prove any equalities, let us try to simplify the crucial term $\vec{x}^\top A\vec{x}$; the intuition behind this is that it looks like the easiest term to use the orthonormal diagonalization on and achieve results.

Let $A = U\Lambda U^\top$ be an orthonormal diagonalization of A. We have

$$\begin{aligned}\vec{x}^\top A\vec{x} &= \vec{x}^\top U\Lambda U^\top \vec{x} && (2.123)\\ &= (U^\top\vec{x})^\top \Lambda (U^\top\vec{x}) && (2.124)\\ &= \vec{y}^\top\Lambda\vec{y} && (2.125)\\ &= \sum_{i=1}^n \lambda_i\{A\} y_i^2 && (2.126)\end{aligned}$$

with the invertible change of variables $\vec{y} \doteq U^\top\vec{x} \iff \vec{x} = U\vec{y}$. Also we note that this change of variables preserves the norm, i.e.,

$$\|\vec{y}\|_2^2 = \left\|U^\top\vec{x}\right\|_2^2 = \vec{x}^\top UU^\top\vec{x} = \vec{x}^\top\vec{x} = \|\vec{x}\|_2^2. \tag{2.127}$$

We now turn to the first equality chain (with $\max$). Immediately, we have

$$\begin{aligned}\max_{\substack{\vec{x}\in\mathbb{R}^n\\ \vec{x}\neq\vec{0}}} \frac{\vec{x}^\top A\vec{x}}{\vec{x}^\top\vec{x}} &= \max_{\substack{\vec{x}\in\mathbb{R}^n\\ \vec{x}\neq\vec{0}}} \frac{\vec{x}^\top A\vec{x}}{\|\vec{x}\|_2^2} && (2.128)\\ &= \max_{\substack{\vec{x}\in\mathbb{R}^n\\ \vec{x}\neq\vec{0}}} \left(\frac{\vec{x}}{\|\vec{x}\|_2}\right)^\top A\left(\frac{\vec{x}}{\|\vec{x}\|_2}\right). && (2.129)\end{aligned}$$

Now, because the norm of our optimization variable $\vec{x}$ does not matter, in that it only affects the objective through its normalization $\vec{x}/\|\vec{x}\|_2$, it is equivalent to optimize over only unit-norm $\vec{x}$, so

$$\max_{\substack{\vec{x}\in\mathbb{R}^n\\ \vec{x}\neq\vec{0}}} \frac{\vec{x}^\top A\vec{x}}{\vec{x}^\top\vec{x}} = \max_{\|\vec{x}\|_2=1} \vec{x}^\top A\vec{x}. \tag{2.130}$$

With the invertible change of variables $\vec{y} = U^\top\vec{x}$ already discussed, we can write

$$\begin{aligned}\max_{\|\vec{x}\|_2=1} \vec{x}^\top A\vec{x} &= \max_{\|\vec{y}\|_2=1} \sum_{i=1}^n \lambda_i y_i^2 && (2.131)\\ &\leq \lambda_{\max}\{A\}\cdot \max_{\|\vec{y}\|_2=1} \sum_{i=1}^n y_i^2 && (2.132)\\ &= \lambda_{\max}\{A\}\cdot \max_{\|\vec{y}\|_2=1} \|\vec{y}\|_2^2 && (2.133)\\ &= \lambda_{\max}\{A\}\cdot \max_{\|\vec{y}\|_2=1} 1 && (2.134)\\ &= \lambda_{\max}\{A\}. && (2.135)\end{aligned}$$

It is left to exhibit a $\vec{y}$ which makes this inequality an equality; indeed, it is achieved when $y_i = 1$ for one i such that $\lambda_i\{A\} = \lambda_{\max}\{A\}$ and $y_i = 0$ otherwise. The achieving $\vec{x}$ can be recovered by $\vec{x} = U\vec{y}$. Note that since this $\vec{y}$ is a standard basis vector $\vec{y} = \vec{e}_i$ for some i such that $\lambda_i\{A\} = \lambda_{\max}\{A\}$, then $\vec{x} = U\vec{y} = U\vec{e}_i = \vec{u}_i$, i.e., the i^{th} column of U, is an eigenvector of A corresponding to the maximum eigenvalue of A.

The analysis for $\lambda_{\min}\{A\}$ goes exactly analogously. □

This characterization motivates defining a new sub-class (or really several new sub-classes) of matrices.

Definition 31 (Positive Semidefinite and Positive Definite Matrices)
Let $A \in \mathbb{S}^n$. We say that A is *positive semidefinite* (PSD), denoted $A \in \mathbb{S}^n_+$, if $\vec{x}^\top A\vec{x} \geq 0$ for all $\vec{x}$. We say that A is *positive definite* (PD), denoted $A \in \mathbb{S}^n_{++}$, if $\vec{x}^\top A\vec{x} > 0$ for all nonzero $\vec{x}$.

There are also negative semidefinite (NSD) and negative definite (ND) symmetric matrices, defined analogously. There are also indefinite symmetric matrices, which are none of the above. It is clear to see that PD matrices are themselves PSD.

Proposition 32
We have $A \in \mathbb{S}^n_+$ if and only if each eigenvalue of A is non-negative. Also, $A \in \mathbb{S}^n_{++}$ if and only if each eigenvalue of A is positive.

Proof. If A is PSD, we have

$$\lambda_{\min}\{A\} = \min_{\substack{\vec{x}\in\mathbb{R}^n \\ \|\vec{x}\|_2=1}} \vec{x}^\top A\vec{x} \geq 0. \tag{2.136}$$

Now suppose that each eigenvalue of A is non-negative. Then

$$0 \leq \lambda_{\min}\{A\} = \min_{\substack{\vec{x}\in\mathbb{R}^n \\ \|\vec{x}\|_2=1}} \vec{x}^\top A\vec{x} \tag{2.137}$$

which implies that $\vec{x}^\top A\vec{x} \geq 0$ for all $\vec{x}$ with unit norm, and by scaling we see that $\vec{x}^\top A\vec{x} \geq 0$ for all $\vec{x} \neq \vec{0}$, while the inequality certainly holds for $\vec{x} = \vec{0}$. Thus $\vec{x}^\top A\vec{x} \geq 0$ for all $\vec{x}$ so $A \in \mathbb{S}^n_+$.

If A is PD, we have

$$\lambda_{\min}\{A\} = \min_{\substack{\vec{x}\in\mathbb{R}^n \\ \|\vec{x}\|_2=1}} \vec{x}^\top A\vec{x} > 0. \tag{2.138}$$

Now suppose that each eigenvalue of A is positive. Then

$$0 < \lambda_{\min}\{A\} = \min_{\substack{\vec{x}\in\mathbb{R}^n \\ \|\vec{x}\|_2=1}} \vec{x}^\top A\vec{x} \tag{2.139}$$

which implies that $\vec{x}^\top A\vec{x} > 0$ for all $\vec{x}$ with unit norm, and by scaling we see that $\vec{x}^\top A\vec{x} > 0$ for all $\vec{x} \neq \vec{0}$, so $A \in \mathbb{S}^n_{++}$. □

The final construction we discuss is that of the *positive semidefinite square root*.

Proposition 33
Let $A \in \mathbb{S}^n_+$. Then there exists a unique symmetric PSD matrix $B \in \mathbb{S}^n_+$, usually denoted $B = A^{1/2}$, such that $A = B^2$.

Proof. Discussion or homework. Note that there are non-symmetric matrices B such that $A = B^2$, but there is a unique PSD B. □

2.5 Principal Component Analysis

Principal components analysis is a way to recover the eponymous principal components of the data. These principal components are those that are most representative of the data structure. Formally, if we have data in $\mathbb{R}^d$, we want to find an underlying p-dimensional linear structure, where $p \ll d$.

This idea has many, many use cases. For example, in modern machine learning, most data has thousands or millions of dimensions. In order to visualize it properly, we need to reduce its dimension to a reasonable number, in order to get an idea about the underlying structure of the data.

Let us first lay out some notation and definitions. Suppose we have the data points $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$. We organize these into a *data matrix* X where data points form the *rows*:[1]

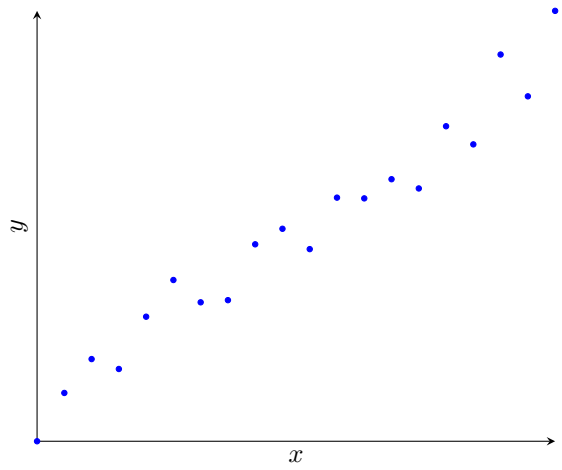
$$X \doteq \begin{bmatrix} \vec{x}_1^\top \\ \vdots \\ \vec{x}_n^\top \end{bmatrix} \in \mathbb{R}^{n \times d} \qquad \text{so that} \qquad X^\top = \begin{bmatrix} \vec{x}_1 & \cdots & \vec{x}_n \end{bmatrix} \in \mathbb{R}^{d \times n}. \tag{2.140}$$

We define the *covariance* matrix $C \in \mathbb{R}^{d \times d}$ by

$$C \doteq \frac{1}{n} X^\top X = \frac{1}{n} \begin{bmatrix} \vec{x}_1 & \cdots & \vec{x}_n \end{bmatrix} \begin{bmatrix} \vec{x}_1^\top \\ \vdots \\ \vec{x}_n^\top \end{bmatrix} = \frac{1}{n} \sum_{i=1}^{n} \vec{x}_i \vec{x}_i^\top. \tag{2.141}$$

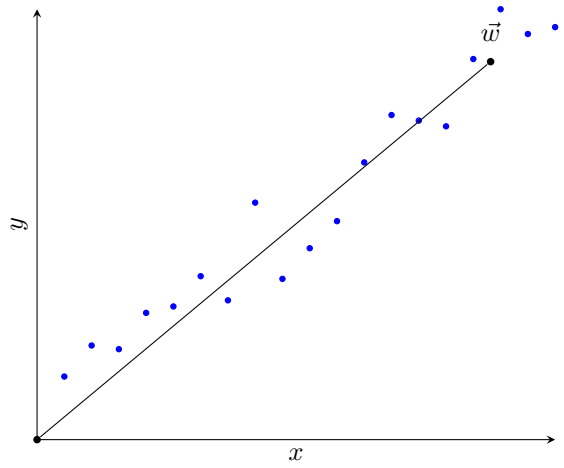
We see that C is symmetric since $X^\top X$ is symmetric, so really $C \in \mathbb{S}^d$.

Before we progress further, let us try to get some intuition for what we might want to be doing. Consider the case $d = 2$ and $p = 1$. Suppose we have the dataset given as below.



While this dataset is clearly fully two-dimensional, there is equally clearly some inherent 1-dimensional linear structure to the data. So when we want to look for an underlying low-dimensional structure, we're looking for something like this. Here, if we could find the direction $\vec{w}$ as in below:

[1] Different textbooks handle this differently. For instance, some textbooks define a data matrix as one where the data points form the columns. If you're unsure, work it out from first principles.



And, if we are in generic $\mathbb{R}^d$ space, we want to find orthonormal vectors $\vec{w}_1, \ldots, \vec{w}_p$ such that projection onto them uncovers the underlying data structure. The process of accurately characterizing these $\vec{w}_i$ is what we will discuss in what follows.

We begin with a motivating example. Consider the MNIST dataset of handwritten digits. Each image is a 28-pixel by 28-pixel grid with each numerical entry in the grid denoting the greyscale value at that grid point. This can be represented by a 28×28 size matrix, or alternatively unrolled into a $28^2 = 784$-dimensional vector. It is impossible to directly visualize 784-dimensional space, so we seek to find $\vec{w}_1, \ldots, \vec{w}_8 \in \mathbb{R}^{784}$ such that the projection onto the $\vec{w}_i$ preserve a lot of structure. Say that we take $\vec{w}_i = \vec{e}_i$, where $\vec{e}_i$ is the i^{th} standard basis vector in $\mathbb{R}^{784}$. Then for most images, the projection onto the $\vec{w}_i$'s will be 0 or near-0. Thus, the projection of all of the data onto the $\vec{w}_i$ preserves almost none of the structure and collapses all points in the dataset to just a few points in $\mathbb{R}^8$. There is instead a much more principled way to choose the $\vec{w}_i$ that will preserve most of the structure.

We now discuss how to choose the first principal component $\vec{w}_1 \in \mathbb{R}^d$. To preserve the structure of the underlying data as much as possible, we want the vectors $\vec{x}_i$ projected onto the span of $\vec{w}_1$ to be as close as possible to the original vectors $\vec{x}_i$. We also want $\|\vec{w}_1\|_2 = 1$. Thus, the error of the projection across all data points is

$$\text{err}(\vec{w}_1) = \frac{1}{n} \sum_{i=1}^{n} \left\| \vec{x}_i - \vec{w}_1(\vec{w}_1^\top \vec{x}_i) \right\|_2^2.$$

Expanding, we have

$$\begin{aligned}
\text{err}(\vec{w}_1) &= \frac{1}{n} \sum_{i=1}^{n} \left\| \vec{x}_i - \vec{w}_1(\vec{w}_1^\top \vec{x}_i) \right\|_2^2 && (2.142)\\
&= \frac{1}{n} \sum_{i=1}^{n} (\vec{x}_i - \vec{w}_1(\vec{w}_1^\top \vec{x}_i))^\top (\vec{x}_i - \vec{w}_1(\vec{w}_1^\top \vec{x}_i)) && (2.143)\\
&= \frac{1}{n} \sum_{i=1}^{n} (\vec{x}_i^\top \vec{x}_i - \vec{x}_i^\top \vec{w}_1(\vec{w}_1^\top \vec{x}_i) - (\vec{w}_1(\vec{w}_1^\top \vec{x}_i))^\top \vec{x}_i + (\vec{w}_1(\vec{w}_1^\top \vec{x}_i))^\top (\vec{w}_1(\vec{w}_1^\top \vec{x}_i))) && (2.144)\\
&= \frac{1}{n} \sum_{i=1}^{n} (\vec{x}_i^\top \vec{x}_i - 2(\vec{x}_i^\top \vec{w}_1)^2 + (\vec{w}_1^\top \vec{w}_1)(\vec{w}_1^\top \vec{x}_i)^2) && (2.145)\\
&= \frac{1}{n} \sum_{i=1}^{n} (\|\vec{x}_i\|_2^2 - 2(\vec{x}_i^\top \vec{w}_1)^2 + (\vec{w}_1^\top \vec{x}_i)^2) && (2.146)\\
&= \frac{1}{n} \sum_{i=1}^{n} (\|\vec{x}_i\|_2^2 - (\vec{x}_i^\top \vec{w}_1)^2). && (2.147)
\end{aligned}$$

Now solving the principal components optimization problem gives

$$\min_{\substack{\vec{w}_1\in\mathbb{R}^d\\ \|\vec{w}_1\|_2=1}} \text{err}(\vec{w}_1) = \min_{\substack{\vec{w}_1\in\mathbb{R}^d\\ \|\vec{w}_1\|_2=1}} \frac{1}{n}\sum_{i=1}^n (\|\vec{x}_i\|_2^2 - (\vec{x}_i^\top \vec{w}_1)^2) \tag{2.148}$$

$$= \frac{1}{n}\sum_{i=1}^n \|\vec{x}_i\|_2^2 + \min_{\substack{\vec{w}_1\in\mathbb{R}^d\\ \|\vec{w}_1\|_2=1}} \frac{1}{n}\sum_{i=1}^n (-(\vec{x}_i^\top \vec{w}_1)^2) \tag{2.149}$$

$$= \frac{1}{n}\sum_{i=1}^n \|\vec{x}_i\|_2^2 - \max_{\substack{\vec{w}_1\in\mathbb{R}^d\\ \|\vec{w}_1\|_2=1}} \frac{1}{n}\sum_{i=1}^n (\vec{x}_i^\top \vec{w}_1)^2 \tag{2.150}$$

$$= \frac{1}{n}\sum_{i=1}^n \|\vec{x}_i\|_2^2 - \max_{\substack{\vec{w}_1\in\mathbb{R}^d\\ \|\vec{w}_1\|_2=1}} \frac{1}{n}\sum_{i=1}^n \vec{w}_1^\top \vec{x}_i\vec{x}_i^\top \vec{w}_1 \tag{2.151}$$

$$= \frac{1}{n}\sum_{i=1}^n \|\vec{x}_i\|_2^2 - \max_{\substack{\vec{w}_1\in\mathbb{R}^d\\ \|\vec{w}_1\|_2=1}} \vec{w}_1^\top \left(\frac{1}{n}\sum_{i=1}^n \vec{x}_i\vec{x}_i^\top\right) \vec{w}_1 \tag{2.152}$$

$$= \frac{1}{n}\sum_{i=1}^n \|\vec{x}_i\|_2^2 - \max_{\substack{\vec{w}_1\in\mathbb{R}^d\\ \|\vec{w}_1\|_2=1}} \vec{w}_1^\top \left(\frac{1}{n}X^\top X\right) \vec{w}_1 \tag{2.153}$$

$$= \frac{1}{n}\sum_{i=1}^n \|\vec{x}_i\|_2^2 - \max_{\substack{\vec{w}_1\in\mathbb{R}^d\\ \|\vec{w}_1\|_2=1}} \vec{w}_1^\top C\vec{w}_1 \tag{2.154}$$

$$= \frac{1}{n}\sum_{i=1}^n \|\vec{x}_i\|_2^2 - \lambda_{\max}\{C\} \tag{2.155}$$

with the $\vec{w}_1$ achieving this upper bound being the eigenvector $\vec{u}_{\max}$ corresponding to the eigenvalue $\lambda_{\max}\{C\}$. Thus, the first principal component is exactly an eigenvector corresponding to the largest eigenvalue of the dot product matrix $C = X^\top X/n$.

This computation is a special case of the singular value decomposition, which is used in practice to compute the PCA of a dataset; understanding this decomposition will allow us to neatly compute the other principal components (i.e., second, third, fourth,...), as well.

2.6 Singular Value Decomposition

Definition 34 (SVD)

Let $A \in \mathbb{R}^{m\times n}$ have rank r. A *singular value decomposition (SVD)* of A is a decomposition of the form

$$A = U\Sigma V^\top \tag{2.156}$$

$$= \begin{bmatrix} U_r & U_{m-r}\end{bmatrix} \begin{bmatrix} \Sigma_r & 0_{r\times(n-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)}\end{bmatrix} \begin{bmatrix} V_r^\top \\ V_{n-r}^\top\end{bmatrix} \tag{2.157}$$

$$= U_r\Sigma_r V_r^\top \tag{2.158}$$

$$= \sum_{i=1}^r \sigma_i \vec{u}_i\vec{v}_i^\top, \tag{2.159}$$

where:

- $U \in \mathbb{R}^{m\times m}$, $U_r \in \mathbb{R}^{m\times r}$, $U_{m-r} \in \mathbb{R}^{m\times(m-r)}$, $V \in \mathbb{R}^{n\times n}$, $V_r \in \mathbb{R}^{n\times r}$, and $V_{n-r} \in \mathbb{R}^{n\times(n-r)}$ are orthonormal matrices, where $U = \begin{bmatrix} U_r & U_{m-r} \end{bmatrix}$ has columns $\vec{u}_1, \ldots, \vec{u}_m$ (*left singular vectors*) and $V = \begin{bmatrix} V_r & V_{n-r} \end{bmatrix}$ has columns $\vec{v}_1, \ldots, \vec{v}_n$ (*right singular vectors*).
- $\Sigma_r = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} \in \mathbb{R}^{r\times r}$ is a diagonal matrix with ordered positive entries $\sigma_1 \geq \cdots \geq \sigma_r > 0$ (*singular values*), and the zero matrices in the $\Sigma = \begin{bmatrix} \Sigma_r & 0_{r\times(n-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{bmatrix}$ matrix are shaped to ensure that $\Sigma \in \mathbb{R}^{m\times n}$.

Suppose that A is tall (so $m > n$) with full column rank n. Then the SVD looks like the following:

$$A = U \begin{bmatrix} \Sigma_n \\ 0_{(m-n)\times n} \end{bmatrix} V^\top. \tag{2.160}$$

On the other hand, if A is wide (so $m < n$) with full row rank m, then the SVD looks like the following:

$$A = U \begin{bmatrix} \Sigma_m & 0_{m\times(n-m)} \end{bmatrix} V^\top. \tag{2.161}$$

The last (summation) form of the SVD is called the *dyadic* SVD; this is because terms of the form $\vec{p}\vec{q}^\top$ are called *dyads*, and the dyadic SVD expresses the matrix A as the sum of dyads.

All forms of the SVD are useful conceptually and computationally, depending on the problem we are working on.

We now discuss a method to construct the SVD. Suppose $A \in \mathbb{R}^{m\times n}$ has rank r. We consider the symmetric matrix $A^\top A$ which has rank r and thus r nonzero eigenvalues, which are positive. We can order its eigenvalues as $\lambda_1 \geq \cdots \geq \lambda_r > \lambda_{r+1} = \cdots = \lambda_n = 0$, say with corresponding orthonormal eigenvectors $\vec{v}_1, \ldots, \vec{v}_n$.

Then, for $i \in \{1, \ldots, r\}$, we define $\sigma_i \doteq \sqrt{\lambda_i} > 0$, and $\vec{u}_i \doteq A\vec{v}_i/\sigma_i$. This only gives us r vectors $\vec{u}_i$, but we need m of them to construct $U \in \mathbb{R}^{m\times m}$. To find the remaining $\vec{u}_i$ we use Gram-Schmidt on the matrix $\begin{bmatrix} \vec{u}_1 & \cdots & \vec{u}_r & I \end{bmatrix} \in \mathbb{R}^{m\times(r+m)}$, throwing out the r vectors whose projection residual onto previously processed vectors is 0.

More formally, we can write an algorithm:

Algorithm 2 Construction of the SVD.

function SVD($A \in \mathbb{R}^{m\times n}$)
  $r \doteq \operatorname{rank}(A)$
  $(\lambda_1, \vec{v}_1), \ldots, (\lambda_n, \vec{v}_n) \doteq$ EIGENPAIRS($A^\top A$) ▷ $\lambda_1 \geq \cdots \geq \lambda_r > \lambda_{r+1} = \cdots = \lambda_n = 0$, and $\vec{v}_i$ orthonormal.
  for $i \in \{1, \ldots, r\}$ **do**
    $\sigma_i \doteq \sqrt{\lambda_i}$ ▷ $\sigma_i > 0$
    $\vec{u}_i \doteq A\vec{v}_i/\sigma_i$
  end for
  $\vec{u}_1, \ldots, \vec{u}_r, \vec{u}_{r+1}, \ldots, \vec{u}_m \doteq$ GRAMSCHMIDT($\begin{bmatrix} \vec{u}_1 & \cdots & \vec{u}_r & I \end{bmatrix}$)
  return $\{\vec{u}_1, \ldots, \vec{u}_m\}, \{\sigma_1, \ldots, \sigma_r\}, \{\vec{v}_1, \ldots, \vec{v}_n\}$
end function

It's clear that Algorithm 2 gives an orthonormal basis $\{\vec{v}_1, \ldots, \vec{v}_n\}$ for $\mathbb{R}^n$ that can be constructed into the orthonormal V matrix, and that it gives singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$. We aim to show two things: the $\{\vec{u}_1, \ldots, \vec{u}_m\}$ are orthonormal, and that $A = U\Sigma V^\top$ where U, Σ, and V are constructed using the returned vectors and scalars.

Proposition 35

In the context of Algorithm 2, $\{\vec{u}_1, \ldots, \vec{u}_m\}$ is an orthonormal set.

Proof. From our invocation of Gram-Schmidt, $\{\vec{u}_{r+1}, \ldots, \vec{u}_m\}$ is an orthonormal set which spans an orthogonal subspace to the span of $\{\vec{u}_1, \ldots, \vec{u}_r\}$. Thus, we need to show that $\{\vec{u}_1, \ldots, \vec{u}_r\}$ are orthonormal.

Indeed, take $1 \leq i < j \leq r$. Then since the $\vec{v}_j$ are orthonormal eigenvectors of $A^\top A$, we have

$$\vec{u}_i^\top \vec{u}_j = \left(\frac{A\vec{v}_i}{\sigma_i}\right)^\top \left(\frac{A\vec{v}_j}{\sigma_j}\right) \tag{2.162}$$

$$= \frac{\vec{v}_i^\top A^\top A \vec{v}_j}{\sigma_i \sigma_j} \tag{2.163}$$

$$= \frac{\lambda_j \vec{v}_i^\top \vec{v}_j}{\sigma_i \sigma_j} \tag{2.164}$$

$$= \frac{\lambda_j}{\sigma_i \sigma_j} \underbrace{\vec{v}_i^\top \vec{v}_j}_{=0} \tag{2.165}$$

$$= 0. \tag{2.166}$$

On the other hand, for a specific $i \in \{1, \ldots, r\}$, using that $\sigma_i^2 = \lambda_i$, we have

$$\|\vec{u}_i\|_2^2 = \left\|\frac{A\vec{v}_i}{\sigma_i}\right\|_2^2 \tag{2.167}$$

$$= \left(\frac{A\vec{v}_i}{\sigma_i}\right)^\top \left(\frac{A\vec{v}_i}{\sigma_i}\right) \tag{2.168}$$

$$= \frac{\vec{v}_i^\top A^\top A \vec{v}_i}{\sigma_i^2} \tag{2.169}$$

$$= \frac{\lambda_i \vec{v}_i^\top \vec{v}_i}{\sigma_i^2} \tag{2.170}$$

$$= \underbrace{\frac{\lambda_i}{\sigma_i^2}}_{=1} \underbrace{\vec{v}_i^\top \vec{v}_i}_{=1} \tag{2.171}$$

$$= 1. \tag{2.172}$$

Thus the set $\{\vec{u}_1, \ldots, \vec{u}_r\}$ is orthonormal, so the whole set $\{\vec{u}_1, \ldots, \vec{u}_m\}$ is orthonormal. □

Proposition 36

In the context of Algorithm 2, we have $A = U\Sigma V^\top$.

Proof. By construction, we have

$$A\vec{v}_i = \sigma_i \vec{u}_i \qquad \text{for all } i \in \{1, \ldots, r\}, \tag{2.173}$$

$$A\vec{v}_i = \vec{0} \qquad \text{for all } i \in \{r+1, \ldots, m\}, \tag{2.174}$$

This gives us

$$AV = A\begin{bmatrix} \vec{v}_1 & \cdots & \vec{v}_r & \vec{v}_{r+1} & \cdots & \vec{v}_n \end{bmatrix} \tag{2.175}$$

$$= \begin{bmatrix} A\vec{v}_1 & \cdots & A\vec{v}_r & A\vec{v}_{r+1} & \cdots & A\vec{v}_n \end{bmatrix} \tag{2.176}$$

$$= \begin{bmatrix} \sigma_1\vec{u}_1 & \cdots & \sigma_r\vec{u}_r & \vec{0} & \cdots & \vec{0} \end{bmatrix} \tag{2.177}$$

$$= \begin{bmatrix} U_r\Sigma_r & 0 \end{bmatrix} \tag{2.178}$$

On the other hand, we have

$$U\Sigma = \begin{bmatrix} U_r & U_{m-r} \end{bmatrix} \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} \tag{2.179}$$

$$= \begin{bmatrix} U_r\Sigma_r + U_{m-r}\cdot 0 & U_r\cdot 0 + U_{m-r}\cdot 0 \end{bmatrix} \tag{2.180}$$

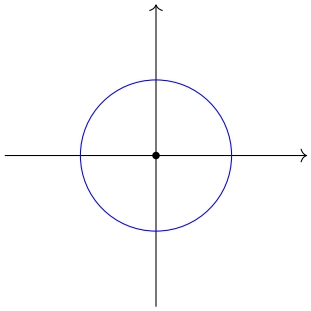
$$= \begin{bmatrix} U_r\Sigma_r & 0 \end{bmatrix}. \tag{2.181}$$

Thus $AV = U\Sigma$. Since V is orthonormal, $V^\top = V^{-1}$, so $A = U\Sigma V^\top$. □

The SVD is not unique: the Gram-Schmidt process could have used any basis for $\mathbb{R}^m$ that wasn't the columns of I and still have been valid; if you had multiple eigenvectors of $A^\top A$ with the same eigenvalue then the choice of eigenvectors in the diagonalization would not be unique; and even if you didn't have multiple eigenvectors with the same eigenvalue, the eigenvectors would only be determined up to a sign change $\vec{v} \mapsto -\vec{v}$ anyways. So there is a lot of ambiguity in the construction, which reflects the non-uniqueness.

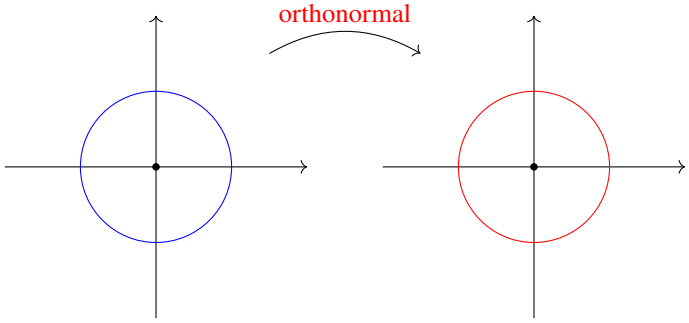
We now discuss the geometry of the SVD, especially how each component of the SVD acts on vectors. To do this we will fix $A \in \mathbb{R}^{2\times 2}$ with SVD $A = U\Sigma V^\top$, and find the behavior of $A\vec{x}$ for all $\vec{x}$ on the unit circle (i.e., with norm 1). We will analyze the behavior of $A\vec{x}$ by using the behavior of $V^\top\vec{x}$, $\Sigma V^\top\vec{x}$ and finally $U\Sigma V^\top\vec{x}$. In the end, we will interpret U as a *rotation or reflection*, Σ as a *scaling*, and $V^\top$ as another *rotation or reflection*.

Before we start, let us discuss what different types of matrices look like as linear transformations. Consider our friendly unit circle:

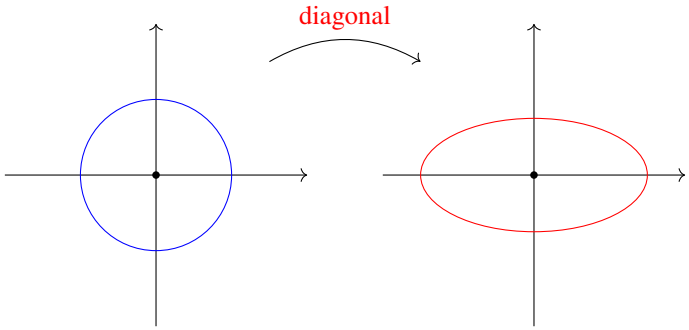


This unit circle consists of all vectors in $\mathbb{R}^2$ with norm 1.

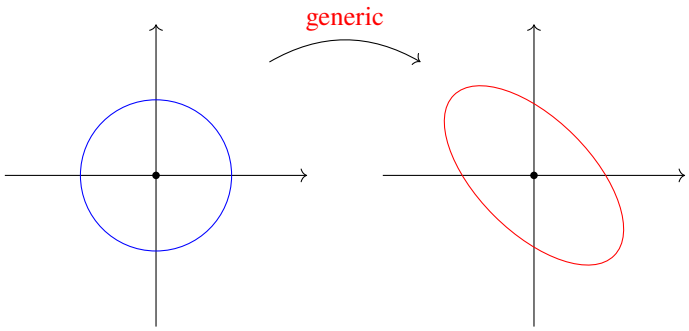
Multiplying each vector in that circle by the same orthonormal matrix will not change the norm of any vector in that circle, and thus will amount to nothing more than a rotation of the circle, thus not changing its shape.



On the other hand, multiplying each vector in the unit circle by the same diagonal matrix will scale the vectors in the coordinate directions. This means that the unit circle will be mapped, in general, to an axis-aligned ellipse.

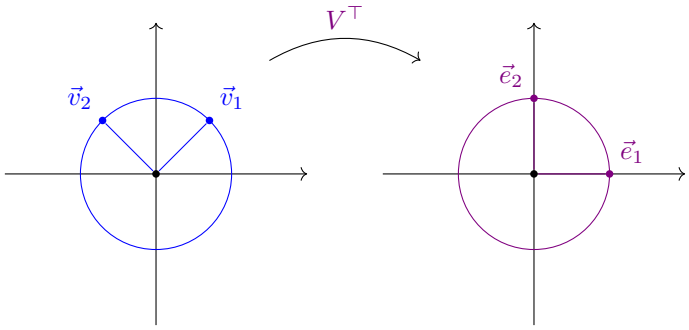


In general, matrices aren't orthonormal or diagonal, and so they will both rotate and scale in various ways. This means that the unit circle will be mapped to an ellipse which isn't necessarily axis-aligned.

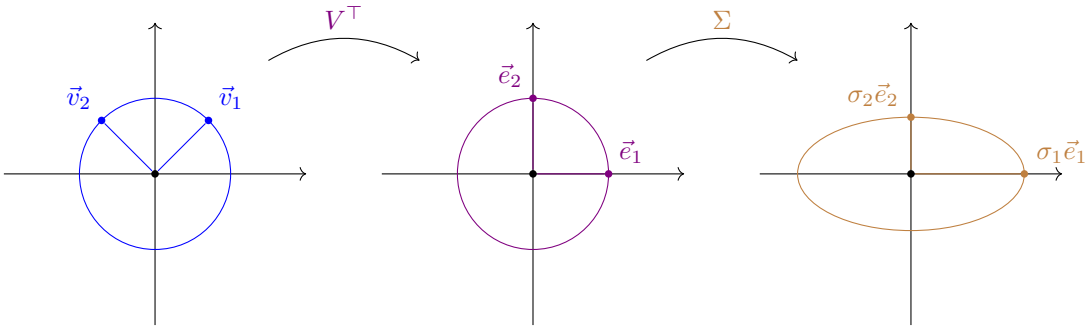


Let us now systematically study such $A = U\Sigma V^\top$ through the lens of the unit circle, as well as where the right singular vectors $\vec{v}_1$ and $\vec{v}_2$ are mapped by A.

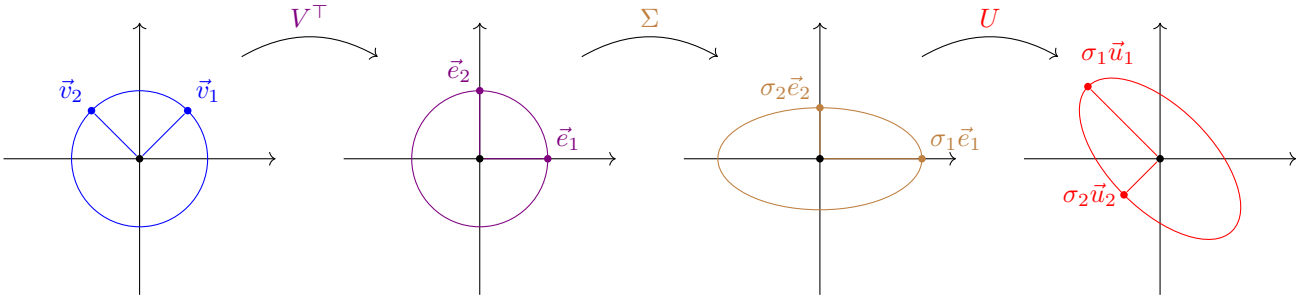
Since $V^\top$ is an orthonormal matrix, it represents a rotation and/or reflection, and so it maps the unit circle to the unit circle, much like our observed figure. Specifically, the right singular vectors $\vec{v}_i$ have $V^\top \vec{v}_i = \vec{e}_i$, so they get mapped onto the standard basis by $V^\top$. This gives the following picture.



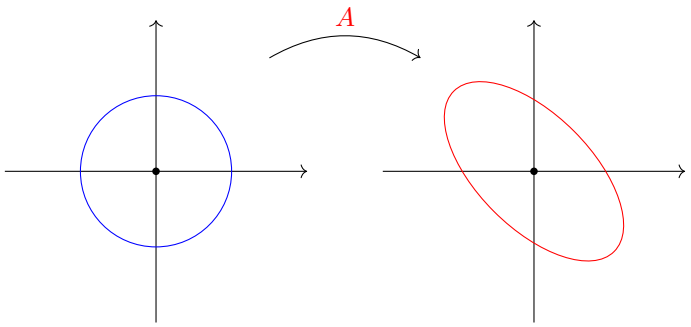
Now, the diagonal matrix Σ will scale each $\vec{e}_i$ by σ_i, obtaining an ellipse.



Finally, the orthonormal matrix U will map this axis-aligned ellipse to an ellipse which isn't necessarily axis-aligned. Specifically, the vectors that we're looking at, i.e., $\sigma_i\vec{e}_i$, have $U(\sigma_i\vec{e}_i) = \sigma_i U\vec{e}_i = \sigma_i\vec{u}_i$. These $\vec{u}_i$ will be the axes of the resulting ellipse in the same sense as $\sigma_i\vec{e}_i$ were the axes of the axis-aligned ellipse.



Recall that we originally started with a depiction that didn't have any fine-grained description of any vectors, yet obtained the same result:



To understand the impact of A on any general vector $\vec{x}$, we write it in the V basis: $\vec{x} = \alpha_1\vec{v}_1 + \alpha_2\vec{v}_2$, and use linearity to obtain $A\vec{x} = \alpha_1\sigma_1\vec{u}_1 + \alpha_2\sigma_2\vec{u}_2$. One can draw this graphically using scaled versions of the above ellipses.

This perspective also says that σ_1 is the maximum scaling of any vector obtained by multiplication by A, and σ_r is the minimum nonzero scaling. (If $r < n$, i.e., A is not full column rank, then there are some nonzero vectors in $\mathbb{R}^n$ which are sent to $\vec{0}$ by A, so the minimum scaling is 0.) You will formally prove this in homework.

2.7 Low-Rank Approximation

Sometimes, in real datasets, matrices have billions or trillions of entries. Storing all of them would be prohibitively expensive, and we would need a way to compress them down into their most important parts. This turns out to be doable via the SVD, as we will see.