20
Linear Algebra

20.1 Vectors

Vectors are a collection of entries (here, we focus only on real numbers). For example, the pair $(1,2)$ is a real vector of size 2, and the 3-tuple $(1,0,2)$ is a real vector of size 3. We primarily categorize vectors by their size. For example, the set of all real vectors of size n is denoted as $\mathbb{R}^n$. Any element of $\mathbb{R}^n$ can be thought of as representing a point (or equivalently, the direction from the origin to the point) in the n-dimensional Cartesian space. A real number in $\mathbb{R}$ is also known as a *scalar*, as opposed to *vectors* in $\mathbb{R}^n$ where $n > 1$.

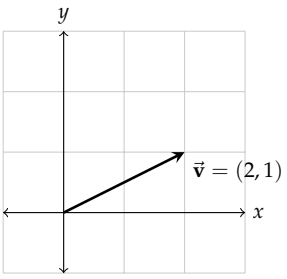

Figure 20.1: A visualization of a vector $\vec{\mathbf{v}} = (2,1)$ in $\mathbb{R}^2$.

20.1.1 Vector Space

We are interested in two operations defined on vectors — vector addition and scalar multiplication. Given vectors $\vec{\mathbf{x}} = (x_1, x_2, \ldots, x_n)$ and $\vec{\mathbf{y}} = (y_1, y_2, \ldots, y_n)$ and a scalar $c \in \mathbb{R}$, the *vector addition* is defined as

$$\vec{\mathbf{x}} + \vec{\mathbf{y}} = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n) \in \mathbb{R}^n$$

where we add each of the coordinates element-wise. As shown in Figure 20.2, vector addition is the process of finding the diagonal of the parallelogram made by the two vectors $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$. The *scalar multiplication* is similarly defined as

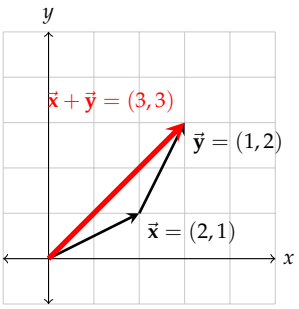

Figure 20.2: A visualization of $\vec{\mathbf{x}} + \vec{\mathbf{y}}$ where $\vec{\mathbf{x}} = (2,1)$ and $\vec{\mathbf{y}} = (1,2)$.

$$c\vec{\mathbf{x}} = (cx_1, cx_2, \ldots, cx_n) \in \mathbb{R}^n$$

As shown in Figure 20.3, scalar multiplication is the process of scaling one vector up or down.

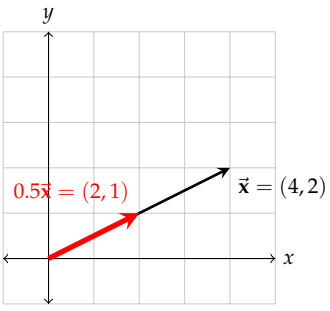

Figure 20.3: A visualization of $0.5\vec{\mathbf{x}}$ where $\vec{\mathbf{x}} = (4,2)$.

$\mathbb{R}^n$ is closed under these two operations — *i.e.*, the resulting vector of either operation is still in $\mathbb{R}^n$. Any subset S of $\mathbb{R}^n$ that is closed under vector addition and scalar multiplication is known as a *subspace* of R^n.

20.1.2 Inner Product

The *inner product* is defined as

$$\vec{\mathbf{x}} \cdot \vec{\mathbf{y}} = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n = \sum_{i=1}^{n} x_i y_i \in \mathbb{R}$$

Closely related to the inner product is the *norm* of a vector, which measures the *length* of it. It is defined as $\|\vec{\mathbf{x}}\| = \sqrt{\vec{\mathbf{x}} \cdot \vec{\mathbf{x}}}$. [1]

[1] There are many other definitions of a norm. This particular one is called an ℓ_2 norm.

Proposition 20.1.1. *The inner product satisfies the following properties:*

- *Symmetry:* $\vec{\mathbf{x}} \cdot \vec{\mathbf{y}} = \vec{\mathbf{y}} \cdot \vec{\mathbf{x}}$
- *Linearity:* $(a_1 \vec{\mathbf{x}}_1 + a_2 \vec{\mathbf{x}}_2) \cdot \vec{\mathbf{y}} = a_1 (\vec{\mathbf{x}}_1 \cdot \vec{\mathbf{y}}) + a_2 (\vec{\mathbf{x}}_2 \cdot \vec{\mathbf{y}})$

and the norm satisfies the following property:

- *Absolute Homogeneity:* $\|a\vec{\mathbf{x}}\| = |a| \, \|\vec{\mathbf{x}}\|$

20.1.3 Linear Independence

Any vector of the form

$$a_1 \vec{\mathbf{x}}_1 + a_2 \vec{\mathbf{x}}_2 + \ldots + a_k \vec{\mathbf{x}}_k$$

where a_i's are scalars and $\vec{\mathbf{x}}_i$'s are vectors is called a *linear combination* of the vectors $\vec{\mathbf{x}}_i$'s. Notice that the zero vector $\vec{\mathbf{0}}$ (*i.e.*, the vector with all zero entries) can always be represented as a linear combination of an arbitrary collection of vectors, if all a_i's are chosen as zero. This is known as a *trivial linear combination*, and any other choice of a_i's is known as a *non-trivial linear combination*.

Definition 20.1.2. *k vectors $\vec{\mathbf{x}}_1, \vec{\mathbf{x}}_2, \ldots, \vec{\mathbf{x}}_k \in \mathbb{R}^n$ are called **linearly dependent** if $\vec{\mathbf{0}}$ can be represented as a non-trivial linear combination of the vectors $\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_k$; or equivalently, if one of the vectors can be represented as a linear combination of the remaining $k-1$ vectors. The vectors that are not linearly dependent with each other are called **linearly independent**.*

Consider the following analogy. Imagine trying to have a family style dinner at a fast food restaurant, where the first person orders a burger, the second person orders a chilli cheese fries, and the third person orders a set menu with a burger and a chili cheese fries. The third person's order did not contribute to the diversity of the food on the dinner table. Similarly, if some set of vectors are linearly dependent, it means that at least one of the vectors is redundant.

Example 20.1.3. *The set $\{(-1,2), (3,0), (1,4)\}$ of three vectors is linearly dependent because*

$$(1,4) = 2 \cdot (-1,2) + (3,0)$$

can be represented as the linear combination of the remaining two vectors.

Example 20.1.4. *The set* $\{(-1,2,1),(3,0,0),(1,4,1)\}$ *of three vectors is linearly independent because there is no way to write one vector as a linear combination of the remaining two vectors.*

20.1.4 Span

Definition 20.1.5. *The* ***span*** *of a set of vectors* $\vec{\mathbf{x}}_1,\ldots,\vec{\mathbf{x}}_k$ *is the set of all vectors that can be represented as a linear combination of* $\vec{\mathbf{x}}_i$*'s.*

Example 20.1.6. $(1,4)$ *is in the span of* $\{(-1,2),(3,0)\}$ *because*

$$(1,4) = 2\cdot(-1,2)+(3,0)$$

Example 20.1.7. $(1,4,1)$ *is not in the span of* $\{(-1,2,1),(3,0,0)\}$ *because there is no way to choose* $a_1,a_2\in\mathbb{R}$ *such that*

$$(1,4,1) = a_1(-1,2,1)+a_2(3,0,0)$$

The span is also known as the *subspace generated by the vectors* $\vec{\mathbf{x}}_1,\ldots,\vec{\mathbf{x}}_k$. This is because if you add any two vectors in the span, or multiply one by a scalar, it is still in the span (*i.e.*, the span is closed under vector addition and scalar multiplication).

Example 20.1.8. *In the* $\mathbb{R}^3$*, the two vectors* $(1,0,0)$ *and* $(0,1,0)$ *span the 2-dimensional XY-plane. Similarly, the vectors* $(1,0,1)$ *and* $(0,2,1)$ *span the 2-dimensional plane* $2x+y-2z=0$*.* [2]

[2] The term *dimension* will be formally defined soon. Here, we rely on your intuition.

In Example 20.1.8, we see examples where 2 vectors span a 2-dimensional subspace. In general, the dimension of the subspace spanned by k vectors can go up to k, but it can also be strictly smaller than k. This is related to the *linear independence* of the vectors.

Proposition 20.1.9. *Given* k *vectors,* $\vec{\mathbf{x}}_1,\ldots,\vec{\mathbf{x}}_k\in\mathbb{R}^n$*, there is a maximum number* $d\geq 1$ *such that there is some subcollection* $\vec{\mathbf{x}}_{i_1},\ldots,\vec{\mathbf{x}}_{i_d}$ *of these vectors that are linearly independent. Then*

$$span(\vec{\mathbf{x}}_1,\ldots,\vec{\mathbf{x}}_k) = span(\vec{\mathbf{x}}_{i_1},\ldots,\vec{\mathbf{x}}_{i_d}) \tag{20.1}$$

is a d*-dimensional subspace of* $\mathbb{R}^n$*.*

Conversely, if we know that the span of the k *vectors is a* d*-dimensional subspace, then the maximum number of vectors that are linearly independent with each other is* d*, and any subcollection of linearly independent* d *vectors satisfies* (20.1)*.*

Proposition 20.1.9 states that the span of some set of k vectors is equivalent to the maximum number d of linearly independent vectors. It also states that the span of the k vectors is equal to the span of the linearly independent d vectors, meaning all of the information

is captured by the d vectors; the remaining $k - d$ vectors are just redundancies. But trying to directly compute the maximum number of linearly independent vectors is inefficient — it may require checking the linear independence of an exponential number of subsets of the vectors. In the next section, we discuss a concept called *matrix rank* that is very closely related to this topic.

20.1.5 Orthogonal Vectors

Definition 20.1.10. *If vectors $\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_k \in \mathbb{R}^n$ satisfy $\vec{\mathbf{x}}_i \cdot \vec{\mathbf{x}}_j = 0$ for any $i \neq j$, then they are called **orthogonal** vectors. In particular, if they also satisfy the condition that $\|\vec{\mathbf{x}}_i\| = 1$ for each i, then they are also **orthonormal**.*

In $\mathbb{R}^n$, orthogonal vectors form a 90 degrees angle with each other.

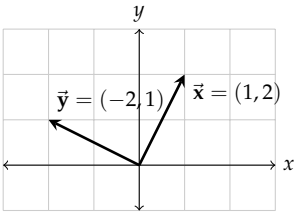

Figure 20.4: A visualization of orthogonal vectors $\vec{\mathbf{x}} = (1, 2)$ and $\vec{\mathbf{y}} = (-2, 1)$.

Example 20.1.11. *The two vectors $(1, 0), (0, 2)$ are orthogonal. So are the vectors $(1, 2), (-2, 1)$.*

Given any set of orthogonal vectors, it is possible to transform it into a set of orthonormal vectors, by normalizing each vector (*i.e.*, scale it such that the norm is 1).

20.1.6 Basis

Definition 20.1.12. *A collection $\{\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_k\}$ of linearly independent vectors in $\mathbb{R}^n$ that span a set S is known as a **basis** of S. In particular, if the vectors of the basis are orthogonal/orthonormal, the basis is called an **orthogonal/orthonormal basis** of S.*

The set S in Definition 20.1.12 can be the entire vector space $\mathbb{R}^n$, but it can also be some subspace of $\mathbb{R}^n$ with a lower dimension.

Example 20.1.13. *The set $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$ of three vectors is a basis for $\mathbb{R}^3$. When we exclude the last vector $(0, 0, 1)$, the set $\{(1, 0, 0), (0, 1, 0)\}$ is a basis of the 2-dimensional XY-plane in $\mathbb{R}^3$.*

Given some subspace S, the basis of S is not unique. However, every basis of S must have the same size — this size is called the *dimension* of S. For a finite dimensional space S, it is known that there exists an *orthogonal* basis of S. There is a well-known algorithm — Gram-Schmidt process — that can transform an arbitrary basis into an orthogonal basis (and eventually an orthonormal basis via normalization).

20.1.7 Projection

Vector projection is the key concept used in the Gram-Schmidt process that computes an orthogonal basis. Given a fixed vector $\vec{\mathbf{a}}$, it decom-

poses any given vector $\vec{\mathbf{x}}$ into a sum of two components — one that is orthogonal to $\vec{\mathbf{a}}$ ("distinct information") and the other that is parallel to $\vec{\mathbf{a}}$ ("redundant information").

Definition 20.1.14 (Vector Projection). *Fix a vector $\vec{\mathbf{a}} \in \mathbb{R}^n$. Given another vector $\vec{\mathbf{x}}$, the **projection of $\vec{\mathbf{x}}$ on $\vec{\mathbf{a}}$** is defined as*

$$proj_{\vec{\mathbf{a}}}(\vec{\mathbf{x}}) = \frac{\vec{\mathbf{x}} \cdot \vec{\mathbf{a}}}{\vec{\mathbf{a}} \cdot \vec{\mathbf{a}}} \vec{\mathbf{a}}$$

and is parallel to the fixed vector $\vec{\mathbf{a}}$. The remaining component

$$\vec{\mathbf{x}} - proj_{\vec{\mathbf{a}}}(\vec{\mathbf{x}})$$

*is called the **rejection of $\vec{\mathbf{x}}$ from $\vec{\mathbf{a}}$** and is orthogonal to $\vec{\mathbf{a}}$.*

Proposition 20.1.15 (Pythagorean Theorem). *If $\vec{\mathbf{x}}, \vec{\mathbf{y}}$ are orthogonal, then*

$$\|\vec{\mathbf{x}} + \vec{\mathbf{y}}\|^2 = \|\vec{\mathbf{x}}\|^2 + \|\vec{\mathbf{y}}\|^2$$

In particular, given two vectors $\vec{\mathbf{a}}, \vec{\mathbf{x}}$, we have

$$\|\vec{\mathbf{x}} - proj_{\vec{\mathbf{a}}}(\vec{\mathbf{x}})\|^2 = \|\vec{\mathbf{x}}\|^2 - \|proj_{\vec{\mathbf{a}}}(\vec{\mathbf{x}})\|^2$$

Now assume we are given a space S and a subspace $T \subset S$. Then a vector $\vec{\mathbf{x}} \in S$ in the larger space does not necessarily belong in T. Instead, we can find a vector $\vec{\mathbf{x}}' \in T$ that is "closest" to $\vec{\mathbf{x}}$ using vector projection. [3]

[3] We ask you to prove this in Problem 7.1.3.

Definition 20.1.16 (Vector Projection on Subspace). *Given a space S, its subspace T with an orthogonal basis $\{\vec{\mathbf{t}}_1, \ldots, \vec{\mathbf{t}}_k\}$, and a vector $\vec{\mathbf{x}} \in S$, the **projection of $\vec{\mathbf{x}}$ on** T is defined as*

$$proj_T(\vec{\mathbf{x}}) = \sum_{i=1}^{k} proj_{\vec{\mathbf{t}}_i}(\vec{\mathbf{x}}) = \sum_{i=1}^{k} \frac{\vec{\mathbf{x}} \cdot \vec{\mathbf{t}}_i}{\vec{\mathbf{t}}_i \cdot \vec{\mathbf{t}}_i} \vec{\mathbf{t}}_i$$

the sum of projection of $\vec{\mathbf{x}}$ on each of the basis vectors of T.

20.2 Matrices

Matrices are a generalization of *vectors* in 2-dimension — a $m \times n$ matrix is a collection of numbers assembled in a rectangular shape of m rows and n columns. The set of all real matrices of size $m \times n$ is denoted as $\mathbb{R}^{m \times n}$. A vector of size n is customarily understood as a column vector — that is, a $n \times 1$ matrix. Also, if $m = n$, then the matrix is known as a *square matrix*.

20.2.1 Matrix Operation

Similarly to vector operations, we are interested in four matrix operations — matrix addition, scalar multiplication, matrix multiplication, and transpose. Given a scalar $c \in \mathbb{R}$ and matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m \times n}$ such that

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{bmatrix} \quad \textit{and} \quad \mathbf{Y} = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,n} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{m,1} & y_{m,2} & \cdots & y_{m,n} \end{bmatrix}$$

the matrix addition is defined as

$$\mathbf{X} + \mathbf{Y} = \begin{bmatrix} x_{1,1} + y_{1,1} & x_{1,2} + y_{1,2} & \cdots & x_{1,n} + y_{1,n} \\ x_{2,1} + y_{2,1} & x_{2,2} + y_{2,2} & \cdots & x_{2,n} + y_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} + y_{m,1} & x_{m,2} + y_{m,2} & \cdots & x_{m,n} + y_{m,n} \end{bmatrix}$$

where we add each of the coordinates element-wise. The *scalar multiplication* is similarly defined as

$$c\mathbf{X} = \begin{bmatrix} cx_{1,1} & cx_{1,2} & \cdots & cx_{1,n} \\ cx_{2,1} & cx_{2,2} & \cdots & cx_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ cx_{m,1} & cx_{m,2} & \cdots & cx_{m,n} \end{bmatrix}$$

The *matrix multiplication* $\mathbf{XY}$ is defined for a matrix $\mathbf{X} \in \mathbb{R}^{\ell \times m}$ and a matrix $\mathbf{Y} \in \mathbb{R}^{m \times n}$; that is, when the number of columns of the first matrix is equal to the number of rows of the second matrix. The output $\mathbf{XY}$ of the matrix multiplication will be a $\ell \times n$ matrix. The (i, j) entry of the matrix $\mathbf{XY}$ is defined as

$$(\mathbf{XY})_{i,j} = \sum_{k=1}^{m} x_{i,k} y_{k,j}$$

That is, it is defined as the inner product of the i-th row of $\mathbf{X}$ and the j-th column of $\mathbf{Y}$.

Proposition 20.2.1. *The above matrix operations satisfy the following properties:*

- $c(\mathbf{XY}) = (c\mathbf{X})\mathbf{Y} = \mathbf{X}(c\mathbf{Y})$
- $(\mathbf{X}_1 + \mathbf{X}_2)\mathbf{Y} = \mathbf{X}_1\mathbf{Y} + \mathbf{X}_2\mathbf{Y}$
- $\mathbf{X}(\mathbf{Y}_1 + \mathbf{Y}_2) = \mathbf{X}\mathbf{Y}_1 + \mathbf{X}\mathbf{Y}_2$

Finally, the *transpose* $\mathbf{X}^\intercal \in \mathbb{R}^{n\times m}$ of a matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$ is the resulting matrix when the entries of $\mathbf{X}$ are reflected down the diagonal. That is,

$$(\mathbf{X}^\intercal)_{i,j} = \mathbf{X}_{j,i}$$

Proposition 20.2.2. *The transpose of a matrix satisfies the following properties:*

- $(\mathbf{X} + \mathbf{Y})^\intercal = \mathbf{X}^\intercal + \mathbf{Y}^\intercal$
- $(c\mathbf{X})^\intercal = c(\mathbf{X}^\intercal)$
- $(\mathbf{XY})^\intercal = \mathbf{Y}^\intercal\mathbf{X}^\intercal$

20.2.2 Matrix and Linear Transformation

Recall that a vector of size n is often considered a $n \times 1$ matrix. Therefore, given a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and a vector $\vec{\mathbf{x}} \in \mathbb{R}^n$, we can define the following operation

$$\vec{\mathbf{y}} = \mathbf{A}\vec{\mathbf{x}} \in \mathbb{R}^m$$

through matrix multiplication. This shows that $\mathbf{A}$ can be understood as a mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$. We see that $a_{i,j}$ (the (i, j) entry of the matrix $\mathbf{A}$) is the coefficient of x_j (the j-th coordinate of the input vector) when computing y_i (the i-th coordinate of the output vector). Since each y_i is linear in terms of each x_j, we say that $\mathbf{A}$ is a *linear transformation*.

20.2.3 Matrix Rank

Matrix rank is one of the most important concepts in basic linear algebra.

Definition 20.2.3. *Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ of m rows and n columns, the number of linearly independent rows is known to be always equal to the number of linearly independent columns. This common number is known as the **rank** of $\mathbf{A}$ and is denoted as $rank(\mathbf{A})$.*

The following property of rank is implied in the definition, but we state it explicitly as follows.

Proposition 20.2.4. *The rank of a matrix is invariant to reordering rows/-columns.*

Example 20.2.5. *Consider the matrix* $M = \begin{bmatrix} 1 & 1 & -2 & 0 \\ -1 & -1 & 2 & 0 \end{bmatrix}$*, we notice that the second row is simply the first row negated, and thus the* rank *of M is* 1.

Example 20.2.6. *Consider the matrix* $M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$*, the rank of* M *is* 3 *because all the row (or column) vectors are linearly independent (they form basis vectors of* $\mathbb{R}^3$*).*

Example 20.2.7. *Consider the matrix* $M = \begin{bmatrix} 1 & 0 & 1 \\ -2 & -3 & 1 \\ 3 & 3 & 0 \end{bmatrix}$*, the rank of* M *is* 2 *because the third row can be expressed as the first row subtracted by the second row.*

When we interpret a matrix as a linear transformation, the rank measures the dimension of the output space.

Proposition 20.2.8. $\mathbf{A} \in \mathbb{R}^{m \times n}$ *has rank* k *if and only if the image of the linear transformation; i.e., the subspace*

$$\{\mathbf{A}\vec{\mathbf{x}} \mid \vec{\mathbf{x}} \in \mathbb{R}^n\}$$

of $\mathbb{R}^m$ *has dimension* k*.*

There are many known algorithms to compute the rank of a matrix. Examples include Gaussian elimination or certain decompositions (expressing a matrix as the product of other matrices with certain properties). Given m vectors in $\mathbb{R}^n$, we can find the maximum number of linearly independent vectors by constructing a matrix with each row equal to each vector [4] and finding the rank of that matrix.

[4] By Proposition 20.2.4, the order of the rows can be arbitrary.

20.2.4 *Eigenvalues and Eigenvectors*

Say we have a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$. This means that the linear transformation expressed by $\mathbf{A}$ is a mapping from $\mathbb{R}^n$ to itself. Most vectors $\vec{\mathbf{x}} \in \mathbb{R}^n$ is mapped to a very "different" vector $\mathbf{A}\vec{\mathbf{x}}$ under this mapping. However, some vectors are "special" and they are mapped to another vector with the same direction.

Definition 20.2.9 (Eigenvalue/Eigenvector)**.** *Given a square matrix* $\mathbf{A} \in \mathbb{R}^{n \times n}$*, if a vector* $\vec{\mathbf{v}} \in \mathbb{R}^n$ *satisfies*

$$\mathbf{A}\vec{\mathbf{v}} = \lambda\vec{\mathbf{v}}$$

for some scalar $\lambda \in \mathbb{R}$*, then* $\vec{\mathbf{v}}$ *is known as an* ***eigenvector*** *of* $\mathbf{A}$*, and* λ *is its corresponding* ***eigenvalue****.*

Each eigenvector can only be associated with one eigenvalue, but each eigenvalue may be associated with multiple eigenvectors.

Proposition 20.2.10. *If* $\vec{\mathbf{x}}, \vec{\mathbf{y}}$ *are both eigenvectors of* $\mathbf{A}$ *for the same eigenvalue* λ*, then any linear combination of them is also an eigenvector for* $\mathbf{A}$ *with the same eigenvalue* λ*.*

Proposition 20.2.10 shows that the set of eigenvectors for a particular eigenvalue forms a subspace, known as the *eigenspace* of that eigenvalue. The dimension of this subspace is known as the *geometric multiplicity* of the eigenvalue. The following result ties together some of the concepts we discussed so far.

Proposition 20.2.11 (Rank-Nullity Theorem)**.** *Given a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, the eigenspace of 0 is the set of all vectors that get mapped to zero vector $\vec{\mathbf{0}}$ under the linear transformation $\mathbf{A}$. This subspace is known as the **null space** of $\mathbf{A}$ and its dimension (i.e., the geometric multiplicity of 0) is known as the **nullity** of $\mathbf{A}$ and is denoted as $nullity(\mathbf{A})$. Then*

$$rank(\mathbf{A}) + nullity(\mathbf{A}) = n$$

20.3 Advanced: SVD/PCA Procedures

Now we briefly introduce a procedure called *Principal Component Analysis (PCA)*, which is commonly used in low-dimensional representation as in Chapter 7.

We are given vectors $\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2, \ldots, \vec{\mathbf{v}}_N \in \mathbb{R}^d$ and a positive integer k and wish to obtain the low-dimensional representation in the sense of Definition 7.1.1 that minimizes ϵ. This is what we mean by "best" representation.

Theorem 20.3.1. *The best low-dimensional representation consists of k eigenvectors corresponding to the top k eigenvalues (largest numerical values) of the matrix $\mathbf{A}\mathbf{A}^\intercal$ where the columns of $\mathbf{A}$ are $\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2, \ldots, \vec{\mathbf{v}}_N$.*

Theorem 20.3.1 shows what the best low-dimensional representation is, but it does not show *how* to compute it. It turns out something called the *Singular Value Decomposition (SVD)* of the matrix $\mathbf{A}$ is useful. It is known that any matrix $\mathbf{A}$ can be decomposed into the following product

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\intercal$$

where $\mathbf{\Sigma}$ is a diagonal matrix with entries equal to the square root of the nonzero eigenvalues of $\mathbf{A}\mathbf{A}^\intercal$ and the columns of $\mathbf{U}$ are the orthonormal eigenvectors of $\mathbf{A}\mathbf{A}^\intercal$, where the i-th column is the eigenvector that corresponds to the eigenvalue at the i-th diagonal entry of $\mathbf{\Sigma}$. There are known computationally efficient algorithms that will perform the SVD of a matrix.

In this section, we will prove Theorem 20.3.1 for the case where $k = 1$. To do this, we need to introduce some preliminary results.

Theorem 20.3.2. *If a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is symmetric (i.e., $\mathbf{A} = \mathbf{A}^\intercal$), then there is an orthonormal basis of $\mathbb{R}^n$ consisting of n eigenvectors of $\mathbf{A}$.* [5]

[5] This is known as the Spectral Theorem.

Proof. A real symmetric matrix is known to be *diagonalizable*, and diagonalizable matrices are known to have n eigenvectors that form a basis for $\mathbb{R}^n$. In particular, the eigenvectors are linearly independent, meaning the eigenvectors corresponding to a particular eigenvalue λ will form a basis for the corresponding eigenspace. Through the Gram-Schmidt process, we can replace some of these eigenvectors such that the eigenvectors for λ are orthogonal to each other. That is, if $\vec{\mathbf{u}}, \vec{\mathbf{v}}$ are eigenvectors for the same eigenvalue λ, then $\vec{\mathbf{u}} \cdot \vec{\mathbf{v}} = 0$. Now assume $\vec{\mathbf{u}}, \vec{\mathbf{v}}$ are two eigenvectors with distinct eigenvalues λ, μ respectively. Then

$$\begin{aligned}\lambda \vec{\mathbf{u}} \cdot \vec{\mathbf{v}} = (\lambda \vec{\mathbf{u}}) \cdot \vec{\mathbf{v}} = (\mathbf{A}\vec{\mathbf{u}}) \cdot \vec{\mathbf{v}} = \sum_{i,j=1}^{n} a_{i,j} u_j v_i \\ = \vec{\mathbf{u}} \cdot (\mathbf{A}^\intercal \vec{\mathbf{v}}) = \vec{\mathbf{u}} \cdot (\mathbf{A}\vec{\mathbf{v}}) = \vec{\mathbf{u}} \cdot (\mu \vec{\mathbf{v}}) = \mu \vec{\mathbf{u}} \cdot \vec{\mathbf{v}}\end{aligned}$$

where the third and the fourth equality can be verified by direct computation. Since $\lambda \neq \mu$, we conclude $\vec{\mathbf{u}} \cdot \vec{\mathbf{v}} = 0$. We have now showed that $\vec{\mathbf{u}} \cdot \vec{\mathbf{v}} = 0$ for any pair of eigenvectors $\vec{\mathbf{u}}, \vec{\mathbf{v}}$ — this means that the basis of eigenvectors is also orthogonal. After normalization, the basis can be made orthonormal. □

The following result is not necessarily needed for the proof of Theorem 20.3.1, but the proofs are similar.

Theorem 20.3.3. *If $\mathbf{A} \in \mathbb{R}^{n \times n}$ is symmetric, then the unit vector $\vec{\mathbf{x}}$ that maximizes $\|\mathbf{A}\vec{\mathbf{x}}\|$ is an eigenvector of $\mathbf{A}$ with an eigenvalue, whose absolute values is the largest out of all eigenvalues.*

Proof. By Theorem 20.3.2, there is an orthonormal basis $\{\vec{\mathbf{u}}_1, \ldots, \vec{\mathbf{u}}_n\}$ of $\mathbb{R}^n$ consisting of eigenvectors of $\mathbf{A}$. Then any vector $\vec{\mathbf{x}}$ is in the span of the eigenvectors and can be represented as the linear combination

$$\vec{\mathbf{x}} = \alpha_1 \vec{\mathbf{u}}_1 + \alpha_2 \vec{\mathbf{u}}_2 + \ldots + \alpha_n \vec{\mathbf{u}}_n$$

for some scalars α_i's. Then

$$\begin{aligned}\|\vec{\mathbf{x}}\|^2 &= \vec{\mathbf{x}} \cdot \vec{\mathbf{x}} \\ &= (\alpha_1 \vec{\mathbf{u}}_1 + \alpha_2 \vec{\mathbf{u}}_2 + \ldots + \alpha_n \vec{\mathbf{u}}_n) \cdot (\alpha_1 \vec{\mathbf{u}}_1 + \alpha_2 \vec{\mathbf{u}}_2 + \ldots + \alpha_n \vec{\mathbf{u}}_n) \\ &= \sum_{i,j=1}^{n} \alpha_i \alpha_j (\vec{\mathbf{u}}_i \cdot \vec{\mathbf{u}}_j) \\ &= \sum_{i=1}^{n} \alpha_i^2\end{aligned}$$

where for the last equality, we use the fact that $\vec{\mathbf{u}}_i$'s are orthonormal — that is, $\vec{\mathbf{u}}_i \cdot \vec{\mathbf{u}}_j = 0$ if $i \neq j$ and $\vec{\mathbf{u}}_i \cdot \vec{\mathbf{u}}_i = 1$. Since $\vec{\mathbf{x}}$ has norm 1, we see

that $\sum_{i=1}^{n} \alpha_i^2 = 1$. Now notice that

$$\begin{aligned}\mathbf{A}\vec{\mathbf{x}} &= \mathbf{A}(\alpha_1\vec{\mathbf{u}}_1 + \alpha_2\vec{\mathbf{u}}_2 + \ldots + \alpha_n\vec{\mathbf{u}}_n) \\ &= \alpha_1\mathbf{A}\vec{\mathbf{u}}_1 + \alpha_2\mathbf{A}\vec{\mathbf{u}}_2 + \ldots + \alpha_n\mathbf{A}\vec{\mathbf{u}}_n \\ &= \alpha_1\lambda_1\vec{\mathbf{u}}_1 + \alpha_2\lambda_2\vec{\mathbf{u}}_2 + \ldots + \alpha_n\lambda_n\vec{\mathbf{u}}_n\end{aligned}$$

where λ_i is the eigenvalue for the eigenvector $\vec{\mathbf{u}}_i$. Following a similar computation as above,

$$\|\mathbf{A}\vec{\mathbf{x}}\|^2 = \sum_{i=1}^{n} \alpha_i^2\lambda_i^2$$

The allocation of weights α_i that will maximize $\sum_{i=1}^{n} \alpha_i^2\lambda_i^2$ while maintaining $\sum_{i=1}^{n} \alpha_i^2 = 1$ is assigning $\alpha_i = \pm 1$ to the eigenvalue λ_i that has the highest value of λ_i^2. This shows that the unit vector $\vec{\mathbf{x}} = \pm\vec{\mathbf{u}}_i$ is an eigenvector with the eigenvalue λ_i. □

We now prove one last preliminary result.

Theorem 20.3.4. *For a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, the matrix $\mathbf{A}\mathbf{A}^\intercal$ is symmetric and its eigenvalues are non-negative.*

Proof. The first part can be verified easily by observing that

$$(\mathbf{A}\mathbf{A}^\intercal)^\intercal = (\mathbf{A}^\intercal)^\intercal\mathbf{A}^\intercal = \mathbf{A}\mathbf{A}^\intercal$$

Now assume $\vec{\mathbf{x}}$ is an eigenvector of $\mathbf{A}$ with eigenvalue λ. Then

$$\mathbf{A}\mathbf{A}^\intercal\vec{\mathbf{x}} = \lambda\vec{\mathbf{x}}$$

We multiply $\vec{\mathbf{x}}^\intercal$ on the left on both sides of the equation.

$$\vec{\mathbf{x}}^\intercal\mathbf{A}\mathbf{A}^\intercal\vec{\mathbf{x}} = \vec{\mathbf{x}}^\intercal(\lambda\vec{\mathbf{x}}) = \lambda\|\vec{\mathbf{x}}\|^2$$

At the same time, notice that

$$\vec{\mathbf{x}}^\intercal\mathbf{A}\mathbf{A}^\intercal\vec{\mathbf{x}} = (\mathbf{A}^\intercal\vec{\mathbf{x}})^\intercal(\mathbf{A}^\intercal\vec{\mathbf{x}}) = \|\mathbf{A}^\intercal\vec{\mathbf{x}}\|^2$$

which shows that

$$\lambda\|\vec{\mathbf{x}}\|^2 = \|\mathbf{A}^\intercal\vec{\mathbf{x}}\|^2$$

Since $\|\vec{\mathbf{x}}\|^2, \|\mathbf{A}^\intercal\vec{\mathbf{x}}\|^2$ are both non-negative, λ is also non-negative. □

We are now ready to (partially) prove the main result of this section.

Proof of Theorem 20.3.1. We prove the case where $k = 1$. Recall that we want to find a vector $\vec{\mathbf{u}}$ that minimizes the error of the low-dimensional representation:

$$\sum_{i=1}^{N} \left\| \vec{\mathbf{v}}_i - \widehat{\vec{\mathbf{v}}}_i \right\|^2$$

where $\widehat{\vec{\mathbf{v}}}_i$ is the low-dimensional representation of $\vec{\mathbf{v}}_i$ that can be computed as

$$\widehat{\vec{\mathbf{v}}}_i = (\vec{\mathbf{v}}_i \cdot \vec{\mathbf{u}})\vec{\mathbf{u}}$$

by the result of Problem 7.1.3. Now by Proposition 20.1.15, we see that

$$\begin{aligned}\sum_{i=1}^{N} \|\vec{\mathbf{v}}_i - (\vec{\mathbf{v}}_i \cdot \vec{\mathbf{u}})\vec{\mathbf{u}}\|^2 &= \sum_{i=1}^{N} \left(\|\vec{\mathbf{v}}_i\|^2 - \|(\vec{\mathbf{v}}_i \cdot \vec{\mathbf{u}})\vec{\mathbf{u}}\|^2 \right) \\ &= \sum_{i=1}^{N} \left(\|\vec{\mathbf{v}}_i\|^2 - (\vec{\mathbf{v}}_i \cdot \vec{\mathbf{u}})^2 \right)\end{aligned}$$

Since we are already given a fixed set of vectors $\vec{\mathbf{v}}_i$, we cannot change the values of $\|\vec{\mathbf{v}}_i\|^2$. Therefore, minimizing the last term of the equation above amounts to maximizing $\sum_{i=1}^{N} (\vec{\mathbf{v}}_i \cdot \vec{\mathbf{u}})^2$. Notice that

$$\sum_{i=1}^{N} (\vec{\mathbf{v}}_i \cdot \vec{\mathbf{u}})^2 = \|\mathbf{A}^\intercal \vec{\mathbf{u}}\|^2 = \vec{\mathbf{u}}^\intercal \mathbf{A}\mathbf{A}^\intercal \vec{\mathbf{u}}$$

By Theorem 20.3.2 and by Theorem 20.3.4, there is an orthonormal basis $\{\vec{\mathbf{u}}_1, \ldots, \vec{\mathbf{u}}_n\}$ of $\mathbb{R}^n$ that consist of the eigenvectors of the matrix $\mathbf{A}\mathbf{A}^\intercal$. Let λ_i be the eigenvalue corresponding to the eigenvector $\vec{\mathbf{u}}_i$. Then similarly to the proof of Theorem 20.3.3, we can represent any vector $\vec{\mathbf{u}}$ as a linear combination of the eigenvectors as

$$\vec{\mathbf{u}} = \alpha_1 \vec{\mathbf{u}}_1 + \alpha_2 \vec{\mathbf{u}}_2 + \ldots + \alpha_n \vec{\mathbf{u}}_n$$

Then we have $\sum_{i=1}^{n} \alpha_i^2 = 1$ and

$$\begin{aligned}\vec{\mathbf{u}}^\intercal \mathbf{A}\mathbf{A}^\intercal \vec{\mathbf{u}} &= (\alpha_1 \vec{\mathbf{u}}_1 + \alpha_2 \vec{\mathbf{u}}_2 + \ldots + \alpha_n \vec{\mathbf{u}}_n)^\intercal \mathbf{A}\mathbf{A}^\intercal (\alpha_1 \vec{\mathbf{u}}_1 + \alpha_2 \vec{\mathbf{u}}_2 + \ldots + \alpha_n \vec{\mathbf{u}}_n) \\ &= (\alpha_1 \vec{\mathbf{u}}_1 + \alpha_2 \vec{\mathbf{u}}_2 + \ldots + \alpha_n \vec{\mathbf{u}}_n)^\intercal (\alpha_1 \lambda_1 \vec{\mathbf{u}}_1 + \alpha_2 \lambda_2 \vec{\mathbf{u}}_2 + \ldots + \alpha_n \lambda_n \vec{\mathbf{u}}_n) \\ &= \sum_{i,j=1}^{n} \alpha_i \alpha_j \lambda_j (\vec{\mathbf{u}}_i \cdot \vec{\mathbf{u}}_j) \\ &= \sum_{i=1}^{n} \alpha_i^2 \lambda_i\end{aligned}$$

Again, the allocation of α_i's that maximize $\sum_{i=1}^{n} \alpha_i^2 \lambda_i$ while maintaining $\sum_{i=1}^{n} \alpha_i^2 = 1$ is assigning $\alpha_i = \pm 1$ to the eigenvector corresponding to the highest value of λ_i. □