3 Linear Algebra

In this section we present important classes of spaces in which our data will live and our operations will take place: vector spaces, metric spaces, normed spaces, and inner product spaces. Generally speaking, these are defined in such a way as to capture one or more important properties of Euclidean space but in a more general way.

3.1 Vector spaces

Vector spaces are the basic setting in which linear algebra happens. A vector space V is a set (the elements of which are called **vectors**) on which two operations are defined: vectors can be added together, and vectors can be multiplied by real numbers[1] called **scalars**. V must satisfy

(i) There exists an additive identity (written $\mathbf{0}$) in V such that $\mathbf{x} + \mathbf{0} = \mathbf{x}$ for all $\mathbf{x} \in V$

(ii) For each $\mathbf{x} \in V$, there exists an additive inverse (written $-\mathbf{x}$) such that $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$

(iii) There exists a multiplicative identity (written 1) in $\mathbb{R}$ such that $1\mathbf{x} = \mathbf{x}$ for all $\mathbf{x} \in V$

(iv) Commutativity: $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$ for all $\mathbf{x}, \mathbf{y} \in V$

(v) Associativity: $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$ and $\alpha(\beta\mathbf{x}) = (\alpha\beta)\mathbf{x}$ for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ and $\alpha, \beta \in \mathbb{R}$

(vi) Distributivity: $\alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \alpha\mathbf{y}$ and $(\alpha + \beta)\mathbf{x} = \alpha\mathbf{x} + \beta\mathbf{x}$ for all $\mathbf{x}, \mathbf{y} \in V$ and $\alpha, \beta \in \mathbb{R}$

A set of vectors $\mathbf{v}_1, \dots, \mathbf{v}_n \in V$ is said to be **linearly independent** if

$$\alpha_1\mathbf{v}_1 + \cdots + \alpha_n\mathbf{v}_n = \mathbf{0} \quad \text{implies} \quad \alpha_1 = \cdots = \alpha_n = 0.$$

The **span** of $\mathbf{v}_1, \dots, \mathbf{v}_n \in V$ is the set of all vectors that can be expressed of a linear combination of them:

$$\operatorname{span}\{\mathbf{v}_1, \dots, \mathbf{v}_n\} = \{\mathbf{v} \in V : \exists \alpha_1, \dots, \alpha_n \text{ such that } \alpha_1\mathbf{v}_1 + \cdots + \alpha_n\mathbf{v}_n = \mathbf{v}\}$$

If a set of vectors is linearly independent and its span is the whole of V, those vectors are said to be a **basis** for V. In fact, every linearly independent set of vectors forms a basis for its span.

If a vector space is spanned by a finite number of vectors, it is said to be **finite-dimensional**. Otherwise it is **infinite-dimensional**. The number of vectors in a basis for a finite-dimensional vector space V is called the **dimension** of V and denoted $\dim V$.

3.1.1 Euclidean space

The quintessential vector space is **Euclidean space**, which we denote $\mathbb{R}^n$. The vectors in this space consist of n-tuples of real numbers:

$$\mathbf{x} = (x_1, x_2, \dots, x_n)$$

For our purposes, it will be useful to think of them as $n \times 1$ matrices, or **column vectors**:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

[1] More generally, vector spaces can be defined over any **field** $\mathbb{F}$. We take $\mathbb{F} = \mathbb{R}$ in this document to avoid an unnecessary diversion into abstract algebra.

Addition and scalar multiplication are defined component-wise on vectors in $\mathbb{R}^n$:

$$\mathbf{x}+\mathbf{y} = \begin{bmatrix} x_1+y_1 \\ \vdots \\ x_n+y_n \end{bmatrix}, \qquad \alpha\mathbf{x} = \begin{bmatrix} \alpha x_1 \\ \vdots \\ \alpha x_n \end{bmatrix}$$

Euclidean space is used to mathematically represent physical space, with notions such as distance, length, and angles. Although it becomes hard to visualize for $n > 3$, these concepts generalize mathematically in obvious ways. Even when you're working in more general settings than $\mathbb{R}^n$, it is often useful to visualize vector addition and scalar multiplication in terms of 2D vectors in the plane or 3D vectors in space.

3.1.2 Subspaces

Vector spaces can contain other vector spaces. If V is a vector space, then $S \subseteq V$ is said to be a **subspace** of V if

(i) $\mathbf{0} \in S$

(ii) S is closed under addition: $\mathbf{x}, \mathbf{y} \in S$ implies $\mathbf{x}+\mathbf{y} \in S$

(iii) S is closed under scalar multiplication: $\mathbf{x} \in S, \alpha \in \mathbb{R}$ implies $\alpha\mathbf{x} \in S$

Note that V is always a subspace of V, as is the trivial vector space which contains only $\mathbf{0}$.

As a concrete example, a line passing through the origin is a subspace of Euclidean space.

If U and W are subspaces of V, then their sum is defined as

$$U + W = \{\mathbf{u}+\mathbf{w} \mid \mathbf{u} \in U, \mathbf{w} \in W\}$$

It is straightforward to verify that this set is also a subspace of V. If $U \cap W = \{\mathbf{0}\}$, the sum is said to be a **direct sum** and written $U \oplus W$. Every vector in $U \oplus W$ can be written uniquely as $\mathbf{u}+\mathbf{w}$ for some $\mathbf{u} \in U$ and $\mathbf{w} \in W$. (This is both a necessary and sufficient condition for a direct sum.)

The dimensions of sums of subspaces obey a friendly relationship (see [4] for proof):

$$\dim(U+W) = \dim U + \dim W - \dim(U \cap W)$$

It follows that

$$\dim(U \oplus W) = \dim U + \dim W$$

since $\dim(U \cap W) = \dim(\{\mathbf{0}\}) = 0$ if the sum is direct.

3.2 Linear maps

A **linear map** is a function $T : V \to W$, where V and W are vector spaces, that satisfies

(i) $T(\mathbf{x}+\mathbf{y}) = T\mathbf{x} + T\mathbf{y}$ for all $\mathbf{x}, \mathbf{y} \in V$

(ii) $T(\alpha\mathbf{x}) = \alpha T\mathbf{x}$ for all $\mathbf{x} \in V, \alpha \in \mathbb{R}$

The standard notational convention for linear maps (which we follow here) is to drop unnecessary parentheses, writing $T\mathbf{x}$ rather than $T(\mathbf{x})$ if there is no risk of ambiguity, and denote composition of linear maps by ST rather than the usual $S \circ T$.

A linear map from V to itself is called a **linear operator**.

Observe that the definition of a linear map is suited to reflect the structure of vector spaces, since it preserves vector spaces' two main operations, addition and scalar multiplication. In algebraic terms, a linear map is called a **homomorphism** of vector spaces. An invertible homomorphism (where the inverse is also a homomorphism) is called an **isomorphism**. If there exists an isomorphism from V to W, then V and W are said to be **isomorphic**, and we write $V \cong W$. Isomorphic vector spaces are essentially "the same" in terms of their algebraic structure. It is an interesting fact that finite-dimensional vector spaces[2] of the same dimension are always isomorphic; if V, W are real vector spaces with $\dim V = \dim W = n$, then we have the natural isomorphism

$$\begin{gathered}\varphi : V \to W \\ \alpha_1\mathbf{v}_1 + \cdots + \alpha_n\mathbf{v}_n \mapsto \alpha_1\mathbf{w}_1 + \cdots + \alpha_n\mathbf{w}_n\end{gathered}$$

where $\mathbf{v}_1, \dots, \mathbf{v}_n$ and $\mathbf{w}_1, \dots, \mathbf{w}_n$ are any bases for V and W. This map is well-defined because every vector in V can be expressed uniquely as a linear combination of $\mathbf{v}_1, \dots, \mathbf{v}_n$. It is straightforward to verify that φ is an isomorphism, so in fact $V \cong W$. In particular, every real n-dimensional vector space is isomorphic to $\mathbb{R}^n$.

3.2.1 The matrix of a linear map

Vector spaces are fairly abstract. To represent and manipulate vectors and linear maps on a computer, we use rectangular arrays of numbers known as **matrices**.

Suppose V and W are finite-dimensional vector spaces with bases $\mathbf{v}_1, \dots, \mathbf{v}_n$ and $\mathbf{w}_1, \dots, \mathbf{w}_m$, respectively, and $T : V \to W$ is a linear map. Then the matrix of T, with entries A_{ij} where $i = 1, \dots, m$, $j = 1, \dots, n$, is defined by

$$T\mathbf{v}_j = A_{1j}\mathbf{w}_1 + \cdots + A_{mj}\mathbf{w}_m$$

That is, the jth column of $\mathbf{A}$ consists of the coordinates of $T\mathbf{v}_j$ in the chosen basis for W.

Conversely, every matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ induces a linear map $T : \mathbb{R}^n \to \mathbb{R}^m$ given by

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

and the matrix of this map with respect to the standard bases of $\mathbb{R}^n$ and $\mathbb{R}^m$ is of course simply $\mathbf{A}$.

If $\mathbf{A} \in \mathbb{R}^{m\times n}$, its **transpose** $\mathbf{A}^\top \in \mathbb{R}^{n\times m}$ is given by $(\mathbf{A}^\top)_{ij} = A_{ji}$ for each (i, j). In other words, the columns of $\mathbf{A}$ become the rows of $\mathbf{A}^\top$, and the rows of $\mathbf{A}$ become the columns of $\mathbf{A}^\top$.

The transpose has several nice algebraic properties that can be easily verified from the definition:

(i) $(\mathbf{A}^\top)^\top = \mathbf{A}$

(ii) $(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$

(iii) $(\alpha\mathbf{A})^\top = \alpha\mathbf{A}^\top$

(iv) $(\mathbf{A}\mathbf{B})^\top = \mathbf{B}^\top\mathbf{A}^\top$

[2] over the same field

3.2.2 Nullspace, range

Some of the most important subspaces are those induced by linear maps. If $T : V \to W$ is a linear map, we define the **nullspace**[3] of T as

$$\text{null}(T) = \{\mathbf{v} \in V \mid T\mathbf{v} = \mathbf{0}\}$$

and the **range** of T as

$$\text{range}(T) = \{\mathbf{w} \in W \mid \exists \mathbf{v} \in V \text{ such that } T\mathbf{v} = \mathbf{w}\}$$

It is a good exercise to verify that the nullspace and range of a linear map are always subspaces of its domain and codomain, respectively.

The **columnspace** of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is the span of its columns (considered as vectors in $\mathbb{R}^m$), and similarly the **rowspace** of $\mathbf{A}$ is the span of its rows (considered as vectors in $\mathbb{R}^n$). It is not hard to see that the columnspace of $\mathbf{A}$ is exactly the range of the linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$ which is induced by $\mathbf{A}$, so we denote it by $\text{range}(\mathbf{A})$ in a slight abuse of notation. Similarly, the rowspace is denoted $\text{range}(\mathbf{A}^\top)$.

It is a remarkable fact that the dimension of the columnspace of $\mathbf{A}$ is the same as the dimension of the rowspace of $\mathbf{A}$. This quantity is called the **rank** of $\mathbf{A}$, and defined as

$$\text{rank}(\mathbf{A}) = \dim \text{range}(\mathbf{A})$$

3.3 Metric spaces

Metrics generalize the notion of distance from Euclidean space (although metric spaces need not be vector spaces).

A **metric** on a set S is a function $d : S \times S \to \mathbb{R}$ that satisfies

(i) $d(x, y) \geq 0$, with equality if and only if $x = y$

(ii) $d(x, y) = d(y, x)$

(iii) $d(x, z) \leq d(x, y) + d(y, z)$ (the so-called **triangle inequality**)

for all $x, y, z \in S$.

A key motivation for metrics is that they allow limits to be defined for mathematical objects other than real numbers. We say that a sequence $\{x_n\} \subseteq S$ converges to the limit x if for any $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that $d(x_n, x) < \epsilon$ for all $n \geq N$. Note that the definition for limits of sequences of real numbers, which you have likely seen in a calculus class, is a special case of this definition when using the metric $d(x, y) = |x - y|$.

3.4 Normed spaces

Norms generalize the notion of length from Euclidean space.

A **norm** on a real vector space V is a function $\|\cdot\| : V \to \mathbb{R}$ that satisfies

[3] It is sometimes called the **kernel** by algebraists, but we eschew this terminology because the word "kernel" has another meaning in machine learning.

(i) $\|\mathbf{x}\| \geq 0$, with equality if and only if $\mathbf{x} = \mathbf{0}$

(ii) $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$

(iii) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (the **triangle inequality** again)

for all $\mathbf{x}, \mathbf{y} \in V$ and all $\alpha \in \mathbb{R}$. A vector space endowed with a norm is called a **normed vector space**, or simply a **normed space**.

Note that any norm on V induces a distance metric on V:

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$$

One can verify that the axioms for metrics are satisfied under this definition and follow directly from the axioms for norms. Therefore any normed space is also a metric space.[4]

We will typically only be concerned with a few specific norms on $\mathbb{R}^n$:

$$\begin{aligned}
\|\mathbf{x}\|_1 &= \sum_{i=1}^{n} |x_i| \\
\|\mathbf{x}\|_2 &= \sqrt{\sum_{i=1}^{n} x_i^2} \\
\|\mathbf{x}\|_p &= \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}} \qquad\qquad (p \geq 1) \\
\|\mathbf{x}\|_\infty &= \max_{1 \leq i \leq n} |x_i|
\end{aligned}$$

Note that the 1- and 2-norms are special cases of the p-norm, and the ∞-norm is the limit of the p-norm as p tends to infinity. We require $p \geq 1$ for the general definition of the p-norm because the triangle inequality fails to hold if $p < 1$. (Try to find a counterexample!)

Here's a fun fact: for any given finite-dimensional vector space V, all norms on V are equivalent in the sense that for two norms $\|\cdot\|_A, \|\cdot\|_B$, there exist constants $\alpha, \beta > 0$ such that

$$\alpha\|\mathbf{x}\|_A \leq \|\mathbf{x}\|_B \leq \beta\|\mathbf{x}\|_A$$

for all $\mathbf{x} \in V$. Therefore convergence in one norm implies convergence in any other norm. This rule may not apply in infinite-dimensional vector spaces such as function spaces, though.

3.5 Inner product spaces

An **inner product** on a real vector space V is a function $\langle\cdot,\cdot\rangle : V \times V \to \mathbb{R}$ satisfying

(i) $\langle\mathbf{x}, \mathbf{x}\rangle \geq 0$, with equality if and only if $\mathbf{x} = \mathbf{0}$

(ii) Linearity in the first slot: $\langle\mathbf{x} + \mathbf{y}, \mathbf{z}\rangle = \langle\mathbf{x}, \mathbf{z}\rangle + \langle\mathbf{y}, \mathbf{z}\rangle$ and $\langle\alpha\mathbf{x}, \mathbf{y}\rangle = \alpha\langle\mathbf{x}, \mathbf{y}\rangle$

(iii) $\langle\mathbf{x}, \mathbf{y}\rangle = \langle\mathbf{y}, \mathbf{x}\rangle$

[4] If a normed space is complete with respect to the distance metric induced by its norm, we say that it is a **Banach space**.

for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ and all $\alpha \in \mathbb{R}$. A vector space endowed with an inner product is called an **inner product space**.

Note that any inner product on V induces a norm on V:

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$$

One can verify that the axioms for norms are satisfied under this definition and follow (almost) directly from the axioms for inner products. Therefore any inner product space is also a normed space (and hence also a metric space).[5]

Two vectors $\mathbf{x}$ and $\mathbf{y}$ are said to be **orthogonal** if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$; we write $\mathbf{x} \perp \mathbf{y}$ for shorthand. Orthogonality generalizes the notion of perpendicularity from Euclidean space. If two orthogonal vectors $\mathbf{x}$ and $\mathbf{y}$ additionally have unit length (i.e. $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$), then they are described as **orthonormal**.

The standard inner product on $\mathbb{R}^n$ is given by

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} x_i y_i = \mathbf{x}^\top \mathbf{y}$$

The matrix notation on the righthand side arises because this inner product is a special case of matrix multiplication where we regard the resulting 1×1 matrix as a scalar. The inner product on $\mathbb{R}^n$ is also often written $\mathbf{x} \cdot \mathbf{y}$ (hence the alternate name **dot product**). The reader can verify that the two-norm $\|\cdot\|_2$ on $\mathbb{R}^n$ is induced by this inner product.

3.5.1 Pythagorean Theorem

The well-known Pythagorean theorem generalizes naturally to arbitrary inner product spaces.

Theorem 1. *If $\mathbf{x} \perp \mathbf{y}$, then*

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

Proof. Suppose $\mathbf{x} \perp \mathbf{y}$, i.e. $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. Then

$$\|\mathbf{x} + \mathbf{y}\|^2 = \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{x} \rangle + \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$$

as claimed. $\square$

3.5.2 Cauchy-Schwarz inequality

This inequality is sometimes useful in proving bounds:

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \, \|\mathbf{y}\|$$

for all $\mathbf{x}, \mathbf{y} \in V$. Equality holds exactly when $\mathbf{x}$ and $\mathbf{y}$ are scalar multiples of each other (or equivalently, when they are linearly dependent).

[5] If an inner product space is complete with respect to the distance metric induced by its inner product, we say that it is a **Hilbert space**.

3.5.3 Orthogonal complements and projections

If $S \subseteq V$ where V is an inner product space, then the **orthogonal complement** of S, denoted $S^\perp$, is the set of all vectors in V that are orthogonal to every element of S:

$$S^\perp = \{\mathbf{v} \in V \mid \mathbf{v} \perp \mathbf{s} \text{ for all } \mathbf{s} \in S\}$$

It is easy to verify that $S^\perp$ is a subspace of V for any $S \subseteq V$. Note that there is no requirement that S itself be a subspace of V. However, if S *is* a (finite-dimensional) subspace of V, we have the following important decomposition.

Proposition 1. *Let V be an inner product space and S be a finite-dimensional subspace of V. Then every $\mathbf{v} \in V$ can be written uniquely in the form*

$$\mathbf{v} = \mathbf{v}_S + \mathbf{v}_\perp$$

where $\mathbf{v}_S \in S$ and $\mathbf{v}_\perp \in S^\perp$.

Proof. Let $\mathbf{u}_1, \dots, \mathbf{u}_m$ be an orthonormal basis for S, and suppose $\mathbf{v} \in V$. Define

$$\mathbf{v}_S = \langle \mathbf{v}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \cdots + \langle \mathbf{v}, \mathbf{u}_m \rangle \mathbf{u}_m$$

and

$$\mathbf{v}_\perp = \mathbf{v} - \mathbf{v}_S$$

It is clear that $\mathbf{v}_S \in S$ since it is in the span of the chosen basis. We also have, for all $i = 1, \dots, m$,

$$\begin{aligned} \langle \mathbf{v}_\perp, \mathbf{u}_i \rangle &= \big\langle \mathbf{v} - (\langle \mathbf{v}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \cdots + \langle \mathbf{v}, \mathbf{u}_m \rangle \mathbf{u}_m), \mathbf{u}_i \big\rangle \\ &= \langle \mathbf{v}, \mathbf{u}_i \rangle - \langle \mathbf{v}, \mathbf{u}_1 \rangle \langle \mathbf{u}_1, \mathbf{u}_i \rangle - \cdots - \langle \mathbf{v}, \mathbf{u}_m \rangle \langle \mathbf{u}_m, \mathbf{u}_i \rangle \\ &= \langle \mathbf{v}, \mathbf{u}_i \rangle - \langle \mathbf{v}, \mathbf{u}_i \rangle \\ &= 0 \end{aligned}$$

which implies $\mathbf{v}_\perp \in S^\perp$.

It remains to show that this decomposition is unique, i.e. doesn't depend on the choice of basis. To this end, let $\mathbf{u}'_1, \dots, \mathbf{u}'_m$ be another orthonormal basis for S, and define $\mathbf{v}'_S$ and $\mathbf{v}'_\perp$ analogously. We claim that $\mathbf{v}'_S = \mathbf{v}_S$ and $\mathbf{v}'_\perp = \mathbf{v}_\perp$.

By definition,

$$\mathbf{v}_S + \mathbf{v}_\perp = \mathbf{v} = \mathbf{v}'_S + \mathbf{v}'_\perp$$

so

$$\underbrace{\mathbf{v}_S - \mathbf{v}'_S}_{\in S} = \underbrace{\mathbf{v}'_\perp - \mathbf{v}_\perp}_{\in S^\perp}$$

From the orthogonality of these subspaces, we have

$$0 = \langle \mathbf{v}_S - \mathbf{v}'_S, \mathbf{v}'_\perp - \mathbf{v}_\perp \rangle = \langle \mathbf{v}_S - \mathbf{v}'_S, \mathbf{v}_S - \mathbf{v}'_S \rangle = \|\mathbf{v}_S - \mathbf{v}'_S\|^2$$

It follows that $\mathbf{v}_S - \mathbf{v}'_S = \mathbf{0}$, i.e. $\mathbf{v}_S = \mathbf{v}'_S$. Then $\mathbf{v}'_\perp = \mathbf{v} - \mathbf{v}'_S = \mathbf{v} - \mathbf{v}_S = \mathbf{v}_\perp$ as well. $\square$

The existence and uniqueness of the decomposition above mean that

$$V = S \oplus S^\perp$$

whenever S is a subspace.

Since the mapping from $\mathbf{v}$ to $\mathbf{v}_S$ in the decomposition above always exists and is unique, we have a well-defined function

$$\begin{aligned} P_S : V &\to S \\ \mathbf{v} &\mapsto \mathbf{v}_S \end{aligned}$$

which is called the **orthogonal projection** onto S. We give the most important properties of this function below.

Proposition 2. *Let S be a finite-dimensional subspace of V. Then*

(i) *For any $\mathbf{v} \in V$ and orthonormal basis $\mathbf{u}_1, \dots, \mathbf{u}_m$ of S,*

$$P_S\mathbf{v} = \langle \mathbf{v}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \cdots + \langle \mathbf{v}, \mathbf{u}_m \rangle \mathbf{u}_m$$

(ii) *For any $\mathbf{v} \in V$, $\mathbf{v} - P_S\mathbf{v} \perp S$.*

(iii) *P_S is a linear map.*

(iv) *P_S is the identity when restricted to S (i.e. $P_S\mathbf{s} = \mathbf{s}$ for all $\mathbf{s} \in S$).*

(v) $\operatorname{range}(P_S) = S$ *and* $\operatorname{null}(P_S) = S^\perp$.

(vi) $P_S^2 = P_S$.

(vii) *For any $\mathbf{v} \in V$, $\|P_S\mathbf{v}\| \leq \|\mathbf{v}\|$.*

(viii) *For any $\mathbf{v} \in V$ and $\mathbf{s} \in S$,*

$$\|\mathbf{v} - P_S\mathbf{v}\| \leq \|\mathbf{v} - \mathbf{s}\|$$

with equality if and only if $\mathbf{s} = P_S\mathbf{v}$. That is,

$$P_S\mathbf{v} = \operatorname*{arg\,min}_{\mathbf{s} \in S} \|\mathbf{v} - \mathbf{s}\|$$

Proof. The first two statements are immediate from the definition of P_S and the work done in the proof of the previous proposition.

In this proof, we abbreviate $P = P_S$ for brevity.

(iii) Suppose $\mathbf{x}, \mathbf{y} \in V$ and $\alpha \in \mathbb{R}$. Write $\mathbf{x} = \mathbf{x}_S + \mathbf{x}_\perp$ and $\mathbf{y} = \mathbf{y}_S + \mathbf{y}_\perp$, where $\mathbf{x}_S, \mathbf{y}_S \in S$ and $\mathbf{x}_\perp, \mathbf{y}_\perp \in S^\perp$. Then

$$\mathbf{x} + \mathbf{y} = \underbrace{\mathbf{x}_S + \mathbf{y}_S}_{\in S} + \underbrace{\mathbf{x}_\perp + \mathbf{y}_\perp}_{\in S^\perp}$$

so $P(\mathbf{x} + \mathbf{y}) = \mathbf{x}_S + \mathbf{y}_S = P\mathbf{x} + P\mathbf{y}$, and

$$\alpha\mathbf{x} = \underbrace{\alpha\mathbf{x}_S}_{\in S} + \underbrace{\alpha\mathbf{x}_\perp}_{\in S^\perp}$$

so $P(\alpha\mathbf{x}) = \alpha\mathbf{x}_S = \alpha P\mathbf{x}$. Thus P is linear.

(iv) If $\mathbf{s} \in S$, then we can write $\mathbf{s} = \mathbf{s} + \mathbf{0}$ where $\mathbf{s} \in S$ and $\mathbf{0} \in S^\perp$, so $P\mathbf{s} = \mathbf{s}$.

(v) range$(P) \subseteq S$: By definition.

range$(P) \supseteq S$: Using the previous result, any $\mathbf{s} \in S$ satisfies $\mathbf{s} = P\mathbf{v}$ for some $\mathbf{v} \in V$ (specifically, $\mathbf{v} = \mathbf{s}$).

null$(P) \subseteq S^\perp$: Suppose $\mathbf{v} \in$ null(P). Write $\mathbf{v} = \mathbf{v}_S + \mathbf{v}_\perp$ where $\mathbf{v}_S \in S$ and $\mathbf{v}_\perp \in S^\perp$. Then $\mathbf{0} = P\mathbf{v} = \mathbf{v}_S$, so $\mathbf{v} = \mathbf{v}_\perp \in S^\perp$.

null$(P) \supseteq S^\perp$: If $\mathbf{v} \in S^\perp$, then $\mathbf{v} = \mathbf{0} + \mathbf{v}$ where $\mathbf{0} \in S$ and $\mathbf{v} \in S^\perp$, so $P\mathbf{v} = \mathbf{0}$.

(vi) For any $\mathbf{v} \in V$,

$$P^2\mathbf{v} = P(P\mathbf{v}) = P\mathbf{v}$$

since $P\mathbf{v} \in S$ and P is the identity on S. Hence $P^2 = P$.

(vii) Suppose $\mathbf{v} \in V$. Then by the Pythagorean theorem,

$$\|\mathbf{v}\|^2 = \|P\mathbf{v} + (\mathbf{v} - P\mathbf{v})\|^2 = \|P\mathbf{v}\|^2 + \|\mathbf{v} - P\mathbf{v}\|^2 \geq \|P\mathbf{v}\|^2$$

The result follows by taking square roots.

(viii) Suppose $\mathbf{v} \in V$ and $\mathbf{s} \in S$. Then by the Pythagorean theorem,-

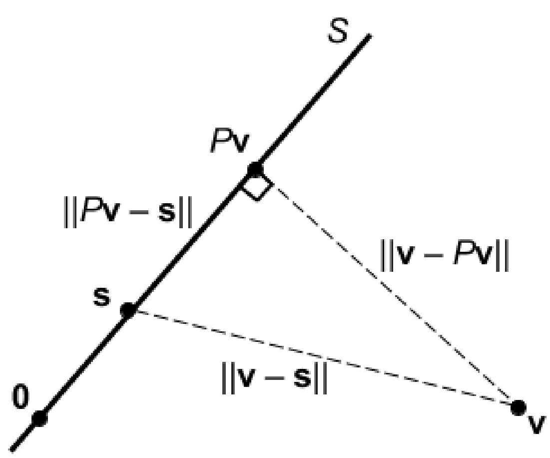
$$\|\mathbf{v} - \mathbf{s}\|^2 = \|(\mathbf{v} - P\mathbf{v}) + (P\mathbf{v} - \mathbf{s})\|^2 = \|\mathbf{v} - P\mathbf{v}\|^2 + \|P\mathbf{v} - \mathbf{s}\|^2 \geq \|\mathbf{v} - P\mathbf{v}\|^2$$

We obtain $\|\mathbf{v} - \mathbf{s}\| \geq \|\mathbf{v} - P\mathbf{v}\|$ by taking square roots. Equality holds iff $\|P\mathbf{v} - \mathbf{s}\|^2 = 0$, which is true iff $P\mathbf{v} = \mathbf{s}$.

$\square$

Any linear map P that satisfies $P^2 = P$ is called a **projection**, so we have shown that P_S is a projection (hence the name).

The last part of the previous result shows that orthogonal projection solves the optimization problem of finding the closest point in S to a given $\mathbf{v} \in V$. This makes intuitive sense from a pictorial representation of the orthogonal projection:



Let us now consider the specific case where S is a subspace of $\mathbb{R}^n$ with orthonormal basis $\mathbf{u}_1, \ldots, \mathbf{u}_m$. Then

$$P_S\mathbf{x} = \sum_{i=1}^m \langle \mathbf{x}, \mathbf{u}_i \rangle \mathbf{u}_i = \sum_{i=1}^m \mathbf{x}^\top \mathbf{u}_i \mathbf{u}_i = \sum_{i=1}^m \mathbf{u}_i \mathbf{u}_i^\top \mathbf{x} = \left(\sum_{i=1}^m \mathbf{u}_i \mathbf{u}_i^\top \right) \mathbf{x}$$

so the operator P_S can be expressed as a matrix

$$\mathbf{P}_S = \sum_{i=1}^{m} \mathbf{u}_i \mathbf{u}_i^\top = \mathbf{U}\mathbf{U}^\top$$

where $\mathbf{U}$ has $\mathbf{u}_1, \dots, \mathbf{u}_m$ as its columns. Here we have used the sum-of-outer-products identity.

3.6 Eigenthings

For a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, there may be vectors which, when $\mathbf{A}$ is applied to them, are simply scaled by some constant. We say that a nonzero vector $\mathbf{x} \in \mathbb{R}^n$ is an **eigenvector** of $\mathbf{A}$ corresponding to **eigenvalue** λ if

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

The zero vector is excluded from this definition because $\mathbf{A}\mathbf{0} = \mathbf{0} = \lambda\mathbf{0}$ for every λ.

We now give some useful results about how eigenvalues change after various manipulations.

Proposition 3. *Let $\mathbf{x}$ be an eigenvector of $\mathbf{A}$ with corresponding eigenvalue λ. Then*

(i) *For any $\gamma \in \mathbb{R}$, $\mathbf{x}$ is an eigenvector of $\mathbf{A} + \gamma\mathbf{I}$ with eigenvalue $\lambda + \gamma$.*

(ii) *If $\mathbf{A}$ is invertible, then $\mathbf{x}$ is an eigenvector of $\mathbf{A}^{-1}$ with eigenvalue λ^{-1}.*

(iii) *$\mathbf{A}^k\mathbf{x} = \lambda^k\mathbf{x}$ for any $k \in \mathbb{Z}$ (where $\mathbf{A}^0 = \mathbf{I}$ by definition).*

Proof. (i) follows readily:

$$(\mathbf{A} + \gamma\mathbf{I})\mathbf{x} = \mathbf{A}\mathbf{x} + \gamma\mathbf{I}\mathbf{x} = \lambda\mathbf{x} + \gamma\mathbf{x} = (\lambda + \gamma)\mathbf{x}$$

(ii) Suppose $\mathbf{A}$ is invertible. Then

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{A}\mathbf{x} = \mathbf{A}^{-1}(\lambda\mathbf{x}) = \lambda\mathbf{A}^{-1}\mathbf{x}$$

Dividing by λ, which is valid because the invertibility of $\mathbf{A}$ implies $\lambda \neq 0$, gives $\lambda^{-1}\mathbf{x} = \mathbf{A}^{-1}\mathbf{x}$.

(iii) The case $k \geq 0$ follows immediately by induction on k. Then the general case $k \in \mathbb{Z}$ follows by combining the $k \geq 0$ case with (ii). $\square$

3.7 Trace

The **trace** of a square matrix is the sum of its diagonal entries:

$$\operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{n} A_{ii}$$

The trace has several nice algebraic properties:

(i) $\operatorname{tr}(\mathbf{A} + \mathbf{B}) = \operatorname{tr}(\mathbf{A}) + \operatorname{tr}(\mathbf{B})$

(ii) $\operatorname{tr}(\alpha\mathbf{A}) = \alpha \operatorname{tr}(\mathbf{A})$

(iii) $\operatorname{tr}(\mathbf{A}^\top) = \operatorname{tr}(\mathbf{A})$

(iv) $\operatorname{tr}(\mathbf{ABCD}) = \operatorname{tr}(\mathbf{BCDA}) = \operatorname{tr}(\mathbf{CDAB}) = \operatorname{tr}(\mathbf{DABC})$

The first three properties follow readily from the definition. The last is known as **invariance under cyclic permutations**. Note that the matrices cannot be reordered arbitrarily, for example $\operatorname{tr}(\mathbf{ABCD}) \neq \operatorname{tr}(\mathbf{BACD})$ in general. Also, there is nothing special about the product of four matrices – analogous rules hold for more or fewer matrices.

Interestingly, the trace of a matrix is equal to the sum of its eigenvalues (repeated according to multiplicity):

$$\operatorname{tr}(\mathbf{A}) = \sum_i \lambda_i(\mathbf{A})$$

3.8 Determinant

The **determinant** of a square matrix can be defined in several different confusing ways, none of which are particularly important for our purposes; go look at an introductory linear algebra text (or Wikipedia) if you need a definition. But it's good to know the properties:

(i) $\det(\mathbf{I}) = 1$

(ii) $\det(\mathbf{A}^\top) = \det(\mathbf{A})$

(iii) $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$

(iv) $\det(\mathbf{A}^{-1}) = \det(\mathbf{A})^{-1}$

(v) $\det(\alpha\mathbf{A}) = \alpha^n \det(\mathbf{A})$

Interestingly, the determinant of a matrix is equal to the product of its eigenvalues (repeated according to multiplicity):

$$\det(\mathbf{A}) = \prod_i \lambda_i(\mathbf{A})$$

3.9 Orthogonal matrices

A matrix $\mathbf{Q} \in \mathbb{R}^{n\times n}$ is said to be **orthogonal** if its columns are pairwise orthonormal. This definition implies that

$$\mathbf{Q}^\top\mathbf{Q} = \mathbf{Q}\mathbf{Q}^\top = \mathbf{I}$$

or equivalently, $\mathbf{Q}^\top = \mathbf{Q}^{-1}$. A nice thing about orthogonal matrices is that they preserve inner products:

$$(\mathbf{Qx})^\top(\mathbf{Qy}) = \mathbf{x}^\top\mathbf{Q}^\top\mathbf{Q}\mathbf{y} = \mathbf{x}^\top\mathbf{I}\mathbf{y} = \mathbf{x}^\top\mathbf{y}$$

A direct result of this fact is that they also preserve 2-norms:

$$\|\mathbf{Qx}\|_2 = \sqrt{(\mathbf{Qx})^\top(\mathbf{Qx})} = \sqrt{\mathbf{x}^\top\mathbf{x}} = \|\mathbf{x}\|_2$$

Therefore multiplication by an orthogonal matrix can be considered as a transformation that preserves length, but may rotate or reflect the vector about the origin.

3.10 Symmetric matrices

A matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is said to be **symmetric** if it is equal to its own transpose ($\mathbf{A} = \mathbf{A}^\top$), meaning that $A_{ij} = A_{ji}$ for all (i, j). This definition seems harmless enough but turns out to have some strong implications. We summarize the most important of these as

Theorem 2. *(Spectral Theorem) If $\mathbf{A} \in \mathbb{R}^{n\times n}$ is symmetric, then there exists an orthonormal basis for $\mathbb{R}^n$ consisting of eigenvectors of $\mathbf{A}$.*

The practical application of this theorem is a particular factorization of symmetric matrices, referred to as the **eigendecomposition** or **spectral decomposition**. Denote the orthonormal basis of eigenvectors $\mathbf{q}_1, \dots, \mathbf{q}_n$ and their eigenvalues $\lambda_1, \dots, \lambda_n$. Let $\mathbf{Q}$ be an orthogonal matrix with $\mathbf{q}_1, \dots, \mathbf{q}_n$ as its columns, and $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$. Since by definition $\mathbf{A}\mathbf{q}_i = \lambda_i \mathbf{q}_i$ for every i, the following relationship holds:

$$\mathbf{AQ} = \mathbf{Q\Lambda}$$

Right-multiplying by $\mathbf{Q}^\top$, we arrive at the decomposition

$$\mathbf{A} = \mathbf{Q\Lambda Q}^\top$$

3.10.1 Rayleigh quotients

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a symmetric matrix. The expression $\mathbf{x}^\top\mathbf{A}\mathbf{x}$ is called a **quadratic form**.

There turns out to be an interesting connection between the quadratic form of a symmetric matrix and its eigenvalues. This connection is provided by the **Rayleigh quotient**

$$R_{\mathbf{A}}(\mathbf{x}) = \frac{\mathbf{x}^\top\mathbf{A}\mathbf{x}}{\mathbf{x}^\top\mathbf{x}}$$

The Rayleigh quotient has a couple of important properties which the reader can (and should!) easily verify from the definition:

(i) **Scale invariance**: for any vector $\mathbf{x} \neq \mathbf{0}$ and any scalar $\alpha \neq 0$, $R_{\mathbf{A}}(\mathbf{x}) = R_{\mathbf{A}}(\alpha\mathbf{x})$.

(ii) If $\mathbf{x}$ is an eigenvector of $\mathbf{A}$ with eigenvalue λ, then $R_{\mathbf{A}}(\mathbf{x}) = \lambda$.

We can further show that the Rayleigh quotient is bounded by the largest and smallest eigenvalues of $\mathbf{A}$. But first we will show a useful special case of the final result.

Proposition 4. *For any $\mathbf{x}$ such that $\|\mathbf{x}\|_2 = 1$,*

$$\lambda_{\min}(\mathbf{A}) \leq \mathbf{x}^\top\mathbf{A}\mathbf{x} \leq \lambda_{\max}(\mathbf{A})$$

with equality if and only if $\mathbf{x}$ is a corresponding eigenvector.

Proof. We show only the max case because the argument for the min case is entirely analogous.

Since $\mathbf{A}$ is symmetric, we can decompose it as $\mathbf{A} = \mathbf{Q\Lambda Q}^\top$. Then use the change of variable $\mathbf{y} = \mathbf{Q}^\top\mathbf{x}$, noting that the relationship between $\mathbf{x}$ and $\mathbf{y}$ is one-to-one and that $\|\mathbf{y}\|_2 = 1$ since $\mathbf{Q}$ is orthogonal. Hence

$$\max_{\|\mathbf{x}\|_2=1} \mathbf{x}^\top\mathbf{A}\mathbf{x} = \max_{\|\mathbf{y}\|_2=1} \mathbf{y}^\top\mathbf{\Lambda}\mathbf{y} = \max_{y_1^2+\cdots+y_n^2=1} \sum_{i=1}^n \lambda_i y_i^2$$

Written this way, it is clear that $\mathbf{y}$ maximizes this expression exactly if and only if it satisfies $\sum_{i\in I} y_i^2 = 1$ where $I = \{i : \lambda_i = \max_{j=1,\dots,n} \lambda_j = \lambda_{\max}(\mathbf{A})\}$ and $y_j = 0$ for $j \notin I$. That is,

I contains the index or indices of the largest eigenvalue. In this case, the maximal value of the expression is

$$\sum_{i=1}^{n} \lambda_i y_i^2 = \sum_{i \in I} \lambda_i y_i^2 = \lambda_{\max}(\mathbf{A}) \sum_{i \in I} y_i^2 = \lambda_{\max}(\mathbf{A})$$

Then writing $\mathbf{q}_1, \dots, \mathbf{q}_n$ for the columns of $\mathbf{Q}$, we have

$$\mathbf{x} = \mathbf{Q}\mathbf{Q}^\top \mathbf{x} = \mathbf{Q}\mathbf{y} = \sum_{i=1}^{n} y_i \mathbf{q}_i = \sum_{i \in I} y_i \mathbf{q}_i$$

where we have used the matrix-vector product identity.

Recall that $\mathbf{q}_1, \dots, \mathbf{q}_n$ are eigenvectors of $\mathbf{A}$ and form an orthonormal basis for $\mathbb{R}^n$. Therefore by construction, the set $\{\mathbf{q}_i : i \in I\}$ forms an orthonormal basis for the eigenspace of $\lambda_{\max}(\mathbf{A})$. Hence $\mathbf{x}$, which is a linear combination of these, lies in that eigenspace and thus is an eigenvector of $\mathbf{A}$ corresponding to $\lambda_{\max}(\mathbf{A})$.

We have shown that $\max_{\|\mathbf{x}\|_2=1} \mathbf{x}^\top \mathbf{A}\mathbf{x} = \lambda_{\max}(\mathbf{A})$, from which we have the general inequality $\mathbf{x}^\top \mathbf{A}\mathbf{x} \leq \lambda_{\max}(\mathbf{A})$ for all unit-length $\mathbf{x}$. $\square$

By the scale invariance of the Rayleigh quotient, we immediately have as a corollary (since $\mathbf{x}^\top \mathbf{A}\mathbf{x} = R_{\mathbf{A}}(\mathbf{x})$ for unit $\mathbf{x}$)

Theorem 3. *(Min-max theorem) For all* $\mathbf{x} \neq \mathbf{0}$*,*

$$\lambda_{\min}(\mathbf{A}) \leq R_{\mathbf{A}}(\mathbf{x}) \leq \lambda_{\max}(\mathbf{A})$$

with equality if and only if $\mathbf{x}$ *is a corresponding eigenvector.*

This is sometimes referred to as a **variational characterization of eigenvalues** because it expresses the smallest/largest eigenvalue of $\mathbf{A}$ in terms of a minimization/maximization problem:

$$\lambda_{\min\,/\,\max}(\mathbf{A}) = \min_{\mathbf{x} \neq \mathbf{0}} / \max_{\mathbf{x} \neq \mathbf{0}} R_{\mathbf{A}}(\mathbf{x})$$

3.11 Positive (semi-)definite matrices

A symmetric matrix $\mathbf{A}$ is **positive semi-definite** if for all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0$. Sometimes people write $\mathbf{A} \succeq 0$ to indicate that $\mathbf{A}$ is positive semi-definite.

A symmetric matrix $\mathbf{A}$ is **positive definite** if for all nonzero $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0$. Sometimes people write $\mathbf{A} \succ 0$ to indicate that $\mathbf{A}$ is positive definite. Note that positive definiteness is a strictly stronger property than positive semi-definiteness, in the sense that every positive definite matrix is positive semi-definite but not vice-versa.

These properties are related to eigenvalues in the following way.

Proposition 5. *A symmetric matrix is positive semi-definite if and only if all of its eigenvalues are nonnegative, and positive definite if and only if all of its eigenvalues are positive.*

Proof. Suppose A is positive semi-definite, and let $\mathbf{x}$ be an eigenvector of $\mathbf{A}$ with eigenvalue λ. Then

$$0 \leq \mathbf{x}^\top \mathbf{A}\mathbf{x} = \mathbf{x}^\top (\lambda \mathbf{x}) = \lambda \mathbf{x}^\top \mathbf{x} = \lambda \|\mathbf{x}\|_2^2$$

Since $\mathbf{x} \neq \mathbf{0}$ (by the assumption that it is an eigenvector), we have $\|\mathbf{x}\|_2^2 > 0$, so we can divide both sides by $\|\mathbf{x}\|_2^2$ to arrive at $\lambda \geq 0$. If $\mathbf{A}$ is positive definite, the inequality above holds strictly, so $\lambda > 0$. This proves one direction.

To simplify the proof of the other direction, we will use the machinery of Rayleigh quotients. Suppose that $\mathbf{A}$ is symmetric and all its eigenvalues are nonnegative. Then for all $\mathbf{x} \neq \mathbf{0}$,

$$0 \leq \lambda_{\min}(\mathbf{A}) \leq R_{\mathbf{A}}(\mathbf{x})$$

Since $\mathbf{x}^\top\mathbf{A}\mathbf{x}$ matches $R_{\mathbf{A}}(\mathbf{x})$ in sign, we conclude that $\mathbf{A}$ is positive semi-definite. If the eigenvalues of $\mathbf{A}$ are all strictly positive, then $0 < \lambda_{\min}(\mathbf{A})$, whence it follows that $\mathbf{A}$ is positive definite. $\square$

As an example of how these matrices arise, consider

Proposition 6. *Suppose* $\mathbf{A} \in \mathbb{R}^{m \times n}$*. Then* $\mathbf{A}^\top\mathbf{A}$ *is positive semi-definite. If* $\operatorname{null}(\mathbf{A}) = \{\mathbf{0}\}$*, then* $\mathbf{A}^\top\mathbf{A}$ *is positive definite.*

Proof. For any $\mathbf{x} \in \mathbb{R}^n$,

$$\mathbf{x}^\top(\mathbf{A}^\top\mathbf{A})\mathbf{x} = (\mathbf{A}\mathbf{x})^\top(\mathbf{A}\mathbf{x}) = \|\mathbf{A}\mathbf{x}\|_2^2 \geq 0$$

so $\mathbf{A}^\top\mathbf{A}$ is positive semi-definite. If $\operatorname{null}(\mathbf{A}) = \{\mathbf{0}\}$, then $\mathbf{A}\mathbf{x} \neq \mathbf{0}$ whenever $\mathbf{x} \neq \mathbf{0}$, so $\|\mathbf{A}\mathbf{x}\|_2^2 > 0$, and thus $\mathbf{A}^\top\mathbf{A}$ is positive definite. $\square$

Positive definite matrices are invertible (since their eigenvalues are nonzero), whereas positive semi-definite matrices might not be. However, if you already have a positive semi-definite matrix, it is possible to perturb its diagonal slightly to produce a positive definite matrix.

Proposition 7. *If* $\mathbf{A}$ *is positive semi-definite and* $\epsilon > 0$*, then* $\mathbf{A} + \epsilon\mathbf{I}$ *is positive definite.*

Proof. Assuming $\mathbf{A}$ is positive semi-definite and $\epsilon > 0$, we have for any $\mathbf{x} \neq \mathbf{0}$ that

$$\mathbf{x}^\top(\mathbf{A} + \epsilon\mathbf{I})\mathbf{x} = \mathbf{x}^\top\mathbf{A}\mathbf{x} + \epsilon\mathbf{x}^\top\mathbf{I}\mathbf{x} = \underbrace{\mathbf{x}^\top\mathbf{A}\mathbf{x}}_{\geq 0} + \underbrace{\epsilon\|\mathbf{x}\|_2^2}_{>0} > 0$$

as claimed. $\square$

An obvious but frequently useful consequence of the two propositions we have just shown is that $\mathbf{A}^\top\mathbf{A} + \epsilon\mathbf{I}$ is positive definite (and in particular, invertible) for *any* matrix $\mathbf{A}$ and any $\epsilon > 0$.

3.11.1 The geometry of positive definite quadratic forms

A useful way to understand quadratic forms is by the geometry of their level sets. A **level set** or **isocontour** of a function is the set of all inputs such that the function applied to those inputs yields a given output. Mathematically, the c-isocontour of f is $\{\mathbf{x} \in \operatorname{dom} f : f(\mathbf{x}) = c\}$.

Let us consider the special case $f(\mathbf{x}) = \mathbf{x}^\top\mathbf{A}\mathbf{x}$ where $\mathbf{A}$ is a positive definite matrix. Since $\mathbf{A}$ is positive definite, it has a unique matrix square root $\mathbf{A}^{\frac{1}{2}} = \mathbf{Q}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Q}^\top$, where $\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$ is the eigendecomposition of $\mathbf{A}$ and $\mathbf{\Lambda}^{\frac{1}{2}} = \operatorname{diag}(\sqrt{\lambda_1}, \dots \sqrt{\lambda_n})$. It is easy to see that this matrix $\mathbf{A}^{\frac{1}{2}}$ is positive definite (consider its eigenvalues) and satisfies $\mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}} = \mathbf{A}$. Fixing a value $c \geq 0$, the c-isocontour of f is the set of $\mathbf{x} \in \mathbb{R}^n$ such that

$$c = \mathbf{x}^\top\mathbf{A}\mathbf{x} = \mathbf{x}^\top\mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}}\mathbf{x} = \|\mathbf{A}^{\frac{1}{2}}\mathbf{x}\|_2^2$$

where we have used the symmetry of $\mathbf{A}^{\frac{1}{2}}$. Making the change of variable $\mathbf{z} = \mathbf{A}^{\frac{1}{2}}\mathbf{x}$, we have the condition $\|\mathbf{z}\|_2 = \sqrt{c}$. That is, the values $\mathbf{z}$ lie on a sphere of radius $\sqrt{c}$. These can be parameterized as $\mathbf{z} = \sqrt{c}\hat{\mathbf{z}}$ where $\hat{\mathbf{z}}$ has $\|\hat{\mathbf{z}}\|_2 = 1$. Then since $\mathbf{A}^{-\frac{1}{2}} = \mathbf{Q}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{Q}^\top$, we have

$$\mathbf{x} = \mathbf{A}^{-\frac{1}{2}}\mathbf{z} = \mathbf{Q}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{Q}^\top\sqrt{c}\hat{\mathbf{z}} = \sqrt{c}\mathbf{Q}\mathbf{\Lambda}^{-\frac{1}{2}}\tilde{\mathbf{z}}$$

where $\tilde{\mathbf{z}} = \mathbf{Q}^\top\hat{\mathbf{z}}$ also satisfies $\|\tilde{\mathbf{z}}\|_2 = 1$ since $\mathbf{Q}$ is orthogonal. Using this parameterization, we see that the solution set $\{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) = c\}$ is the image of the unit sphere $\{\tilde{\mathbf{z}} \in \mathbb{R}^n : \|\tilde{\mathbf{z}}\|_2 = 1\}$ under the invertible linear map $\mathbf{x} = \sqrt{c}\mathbf{Q}\mathbf{\Lambda}^{-\frac{1}{2}}\tilde{\mathbf{z}}$.

What we have gained with all these manipulations is a clear algebraic understanding of the c-isocontour of f in terms of a sequence of linear transformations applied to a well-understood set. We begin with the unit sphere, then scale every axis i by $\lambda_i^{-\frac{1}{2}}$, resulting in an axis-aligned ellipsoid. Observe that the axis lengths of the ellipsoid are proportional to the inverse square roots of the eigenvalues of $\mathbf{A}$. Hence larger eigenvalues correspond to shorter axis lengths, and vice-versa.

Then this axis-aligned ellipsoid undergoes a rigid transformation (i.e. one that preserves length and angles, such as a rotation/reflection) given by $\mathbf{Q}$. The result of this transformation is that the axes of the ellipse are no longer along the coordinate axes in general, but rather along the directions given by the corresponding eigenvectors. To see this, consider the unit vector $\mathbf{e}_i \in \mathbb{R}^n$ that has $[\mathbf{e}_i]_j = \delta_{ij}$. In the pre-transformed space, this vector points along the axis with length proportional to $\lambda_i^{-\frac{1}{2}}$. But after applying the rigid transformation $\mathbf{Q}$, the resulting vector points in the direction of the corresponding eigenvector $\mathbf{q}_i$, since

$$\mathbf{Q}\mathbf{e}_i = \sum_{j=1}^{n}[\mathbf{e}_i]_j\mathbf{q}_j = \mathbf{q}_i$$

where we have used the matrix-vector product identity from earlier.

In summary: the isocontours of $f(\mathbf{x}) = \mathbf{x}^\top\mathbf{A}\mathbf{x}$ are ellipsoids such that the axes point in the directions of the eigenvectors of $\mathbf{A}$, and the radii of these axes are proportional to the inverse square roots of the corresponding eigenvalues.

3.12 Singular value decomposition

Singular value decomposition (SVD) is a widely applicable tool in linear algebra. Its strength stems partially from the fact that *every matrix* $\mathbf{A} \in \mathbb{R}^{m\times n}$ has an SVD (even non-square matrices)! The decomposition goes as follows:

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

where $\mathbf{U} \in \mathbb{R}^{m\times m}$ and $\mathbf{V} \in \mathbb{R}^{n\times n}$ are orthogonal matrices and $\mathbf{\Sigma} \in \mathbb{R}^{m\times n}$ is a diagonal matrix with the **singular values** of $\mathbf{A}$ (denoted σ_i) on its diagonal.

Only the first $r = \operatorname{rank}(\mathbf{A})$ singular values are nonzero, and by convention, they are in non-increasing order, i.e.

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_{\min(m,n)} = 0$$

Another way to write the SVD (cf. the sum-of-outer-products identity) is

$$\mathbf{A} = \sum_{i=1}^{r}\sigma_i\mathbf{u}_i\mathbf{v}_i^\top$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$ are the ith columns of $\mathbf{U}$ and $\mathbf{V}$, respectively.

Observe that the SVD factors provide eigendecompositions for $\mathbf{A}^\top\mathbf{A}$ and $\mathbf{A}\mathbf{A}^\top$:

$$\mathbf{A}^\top\mathbf{A} = (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)^\top\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{V}\mathbf{\Sigma}^\top\mathbf{U}^\top\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{V}\mathbf{\Sigma}^\top\mathbf{\Sigma}\mathbf{V}^\top$$
$$\mathbf{A}\mathbf{A}^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{V}\mathbf{\Sigma}^\top\mathbf{U}^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^\top\mathbf{U}^\top$$

It follows immediately that the columns of $\mathbf{V}$ (the **right-singular vectors** of $\mathbf{A}$) are eigenvectors of $\mathbf{A}^\top\mathbf{A}$, and the columns of $\mathbf{U}$ (the **left-singular vectors** of $\mathbf{A}$) are eigenvectors of $\mathbf{A}\mathbf{A}^\top$.

The matrices $\mathbf{\Sigma}^\top\mathbf{\Sigma}$ and $\mathbf{\Sigma}\mathbf{\Sigma}^\top$ are not necessarily the same size, but both are diagonal with the squared singular values σ_i^2 on the diagonal (plus possibly some zeros). Thus the singular values of $\mathbf{A}$ are the square roots of the eigenvalues of $\mathbf{A}^\top\mathbf{A}$ (or equivalently, of $\mathbf{A}\mathbf{A}^\top$)[6].

3.13 Fundamental Theorem of Linear Algebra

Despite its fancy name, the "Fundamental Theorem of Linear Algebra" is not a universally-agreed-upon theorem; there is some ambiguity as to exactly what statements it includes. The version we present here is sufficient for our purposes.

Theorem 4. *If* $\mathbf{A} \in \mathbb{R}^{m\times n}$*, then*

(i) $\operatorname{null}(\mathbf{A}) = \operatorname{range}(\mathbf{A}^\top)^\perp$

(ii) $\operatorname{null}(\mathbf{A}) \oplus \operatorname{range}(\mathbf{A}^\top) = \mathbb{R}^n$

(iii) $\underbrace{\dim \operatorname{range}(\mathbf{A})}_{\operatorname{rank}(\mathbf{A})} + \dim \operatorname{null}(\mathbf{A}) = n$.[7]

(iv) *If* $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ *is the singular value decomposition of* $\mathbf{A}$*, then the columns of* $\mathbf{U}$ *and* $\mathbf{V}$ *form orthonormal bases for the four "fundamental subspaces" of* $\mathbf{A}$*:*

Subspace	*Columns*
$\operatorname{range}(\mathbf{A})$	*The first r columns of* $\mathbf{U}$
$\operatorname{range}(\mathbf{A}^\top)$	*The first r columns of* $\mathbf{V}$
$\operatorname{null}(\mathbf{A}^\top)$	*The last $m-r$ columns of* $\mathbf{U}$
$\operatorname{null}(\mathbf{A})$	*The last $n-r$ columns of* $\mathbf{V}$

where $r = \operatorname{rank}(\mathbf{A})$.

Proof. (i) Let $\mathbf{a}_1, \dots, \mathbf{a}_m$ denote the rows of $\mathbf{A}$. Then

$$\begin{aligned}
\mathbf{x} \in \operatorname{null}(\mathbf{A}) &\iff \mathbf{A}\mathbf{x} = \mathbf{0} \\
&\iff \mathbf{a}_i^\top\mathbf{x} = 0 \text{ for all } i = 1, \dots, m \\
&\iff (\alpha_1\mathbf{a}_1 + \cdots + \alpha_m\mathbf{a}_m)^\top\mathbf{x} = 0 \text{ for all } \alpha_1, \dots, \alpha_m \\
&\iff \mathbf{v}^\top\mathbf{x} = 0 \text{ for all } \mathbf{v} \in \operatorname{range}(\mathbf{A}^\top) \\
&\iff \mathbf{x} \in \operatorname{range}(\mathbf{A}^\top)^\perp
\end{aligned}$$

which proves the result.

[6] Recall that $\mathbf{A}^\top\mathbf{A}$ and $\mathbf{A}\mathbf{A}^\top$ are positive semi-definite, so their eigenvalues are nonnegative, and thus taking square roots is always well-defined.

[7] This result is sometimes referred to by itself as the **rank-nullity theorem**.

(ii) Recall our previous result on orthogonal complements: if S is a finite-dimensional subspace of V, then $V = S \oplus S^\perp$. Thus the claim follows from the previous part (take $V = \mathbb{R}^n$ and $S = \text{range}(\mathbf{A}^\top)$).

(iii) Recall that if U and W are subspaces of a finite-dimensional vector space V, then $\dim(U \oplus W) = \dim U + \dim W$. Thus the claim follows from the previous part, using the fact that $\dim \text{range}(\mathbf{A}) = \dim \text{range}(\mathbf{A}^\top)$.

□

A direct result of (ii) is that every $\mathbf{x} \in \mathbb{R}^n$ can be written (uniquely) in the form

$$\mathbf{x} = \mathbf{A}^\top \mathbf{v} + \mathbf{w}$$

for some $\mathbf{v} \in \mathbb{R}^m, \mathbf{w} \in \mathbb{R}^n$, where $\mathbf{A}\mathbf{w} = \mathbf{0}$.

Note that there is some asymmetry in the theorem, but analogous statements can be obtained by applying the theorem to $\mathbf{A}^\top$.

3.14 Operator and matrix norms

If V and W are vector spaces, then the set of linear maps from V to W forms another vector space, and the norms defined on V and W induce a norm on this space of linear maps. If $T : V \to W$ is a linear map between normed spaces V and W, then the **operator norm** is defined as

$$\|T\|_{\text{op}} = \max_{\substack{\mathbf{x} \in V \\ \mathbf{x} \neq \mathbf{0}}} \frac{\|T\mathbf{x}\|_W}{\|\mathbf{x}\|_V}$$

An important class of this general definition is when the domain and codomain are $\mathbb{R}^n$ and $\mathbb{R}^m$, and the p-norm is used in both cases. Then for a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, we can define the matrix p-norm

$$\|\mathbf{A}\|_p = \max_{\mathbf{x} \neq \mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p}$$

In the special cases $p = 1, 2, \infty$ we have

$$\begin{aligned}
\|\mathbf{A}\|_1 &= \max_{1 \leq j \leq n} \sum_{i=1}^m |A_{ij}| \\
\|\mathbf{A}\|_\infty &= \max_{1 \leq i \leq m} \sum_{j=1}^n |A_{ij}| \\
\|\mathbf{A}\|_2 &= \sigma_1(\mathbf{A})
\end{aligned}$$

where σ_1 denotes the largest singular value. Note that the induced 1- and ∞-norms are simply the maximum absolute column and row sums, respectively. The induced 2-norm (often called the **spectral norm**) simplifies to σ_1 by the properties of Rayleigh quotients proved earlier; clearly

$$\arg\max_{\mathbf{x} \neq \mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \arg\max_{\mathbf{x} \neq \mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2} = \arg\max_{\mathbf{x} \neq \mathbf{0}} \frac{\mathbf{x}^\top \mathbf{A}^\top \mathbf{A}\mathbf{x}}{\mathbf{x}^\top \mathbf{x}}$$

and we have seen that the rightmost expression is maximized by an eigenvector of $\mathbf{A}^\top\mathbf{A}$ corresponding to its largest eigenvalue, $\lambda_{\max}(\mathbf{A}^\top\mathbf{A}) = \sigma_1^2(\mathbf{A})$.

By definition, these induced matrix norms have the important property that

$$\|\mathbf{A}\mathbf{x}\|_p \leq \|\mathbf{A}\|_p \|\mathbf{x}\|_p$$

for any $\mathbf{x}$. They are also **submultiplicative** in the following sense.

Proposition 8. $\|\mathbf{AB}\|_p \leq \|\mathbf{A}\|_p \|\mathbf{B}\|_p$

Proof. For any $\mathbf{x}$,

$$\|\mathbf{ABx}\|_p \leq \|\mathbf{A}\|_p \|\mathbf{Bx}\|_p \leq \|\mathbf{A}\|_p \|\mathbf{B}\|_p \|\mathbf{x}\|_p$$

so

$$\|\mathbf{AB}\|_p = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{ABx}\|}{\|\mathbf{x}\|_p} \leq \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\|_p \|\mathbf{B}\|_p \|\mathbf{x}\|_p}{\|\mathbf{x}\|_p} = \|\mathbf{A}\|_p \|\mathbf{B}\|_p$$

□

These are not the only matrix norms, however. Another frequently used is the **Frobenius norm**

$$\|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2} = \sqrt{\operatorname{tr}(\mathbf{A}^\top\mathbf{A})} = \sqrt{\sum_{i=1}^{\min(m,n)} \sigma_i^2(\mathbf{A})}$$

The first equivalence follows straightforwardly by expanding the definitions of matrix multiplication and trace. For the second, observe that (writing $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ as before)

$$\operatorname{tr}(\mathbf{A}^\top\mathbf{A}) = \operatorname{tr}(\mathbf{V}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}\mathbf{V}^\top) = \operatorname{tr}(\mathbf{V}^\top\mathbf{V}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}) = \operatorname{tr}(\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}) = \sum_{i=1}^{\min(m,n)} \sigma_i^2(\mathbf{A})$$

using the cyclic property of trace and orthogonality of $\mathbf{V}$.

A matrix norm $\|\cdot\|$ is said to be **unitary invariant** if

$$\|\mathbf{UAV}\| = \|\mathbf{A}\|$$

for all orthogonal $\mathbf{U}$ and $\mathbf{V}$ of appropriate size. Unitary invariant norms essentially depend only on the singular values of a matrix, since for such norms,

$$\|\mathbf{A}\| = \|\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top\| = \|\boldsymbol{\Sigma}\|$$

Two particular norms we have seen, the spectral norm and the Frobenius norm, can be expressed solely in terms of a matrix's singular values.

Proposition 9. *The spectral norm and the Frobenius norm are unitary invariant.*

Proof. For the Frobenius norm, the claim follows from

$$\operatorname{tr}((\mathbf{UAV})^\top\mathbf{UAV}) = \operatorname{tr}(\mathbf{V}^\top\mathbf{A}^\top\mathbf{U}^\top\mathbf{UAV}) = \operatorname{tr}(\mathbf{V}\mathbf{V}^\top\mathbf{A}^\top\mathbf{A}) = \operatorname{tr}(\mathbf{A}^\top\mathbf{A})$$

For the spectral norm, recall that $\|\mathbf{Ux}\|_2 = \|\mathbf{x}\|_2$ for any orthogonal $\mathbf{U}$. Thus

$$\|\mathbf{UAV}\|_2 = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{UAVx}\|_2}{\|\mathbf{x}\|_2} = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{AVx}\|_2}{\|\mathbf{x}\|_2} = \max_{\mathbf{y}\neq\mathbf{0}} \frac{\|\mathbf{Ay}\|_2}{\|\mathbf{y}\|_2} = \|\mathbf{A}\|_2$$

where we have used the change of variable $\mathbf{y} = \mathbf{V}^\top\mathbf{x}$, which satisfies $\|\mathbf{y}\|_2 = \|\mathbf{x}\|_2$. Since $\mathbf{V}^\top$ is invertible, $\mathbf{x}$ and $\mathbf{y}$ are in one-to-one correspondence, and in particular $\mathbf{y} = \mathbf{0}$ if and only if $\mathbf{x} = \mathbf{0}$. Hence maximizing over $\mathbf{y} \neq \mathbf{0}$ is equivalent to maximizing over $\mathbf{x} \neq \mathbf{0}$. □

3.15 Low-rank approximation

An important practical application of the SVD is to compute **low-rank approximations** to matrices. That is, given some matrix, we want to find another matrix of the same dimensions but lower rank such that the two matrices are close as measured by some norm. Such an approximation can be used to reduce the amount of data needed to store a matrix, while retaining most of its information.

A remarkable result known as the **Eckart-Young-Mirsky theorem** tells us that the optimal matrix can be computed easily from the SVD, as long as the norm in question is unitary invariant (e.g., the spectral norm or Frobenius norm).

Theorem 5. *(Eckart-Young-Mirsky) Let $\|\cdot\|$ be a unitary invariant matrix norm. Suppose $\mathbf{A} \in \mathbb{R}^{m \times n}$, where $m \geq n$, has singular value decomposition $\mathbf{A} = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$. Then the best rank-$k$ approximation to $\mathbf{A}$, where $k \leq \operatorname{rank}(\mathbf{A})$, is given by*

$$\mathbf{A}_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$$

in the sense that

$$\|\mathbf{A} - \mathbf{A}_k\| \leq \|\mathbf{A} - \tilde{\mathbf{A}}\|$$

for any $\tilde{\mathbf{A}} \in \mathbb{R}^{m \times n}$ with $\operatorname{rank}(\tilde{\mathbf{A}}) \leq k$.

The proof of the general case requires a fair amount of work, so we prove only the special case where $\|\cdot\|$ is the spectral norm.

Proof. First we compute

$$\|\mathbf{A} - \mathbf{A}_k\|_2 = \left\| \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top - \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top \right\|_2 = \left\| \sum_{i=k+1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top \right\|_2 = \sigma_{k+1}$$

Let $\tilde{\mathbf{A}} \in \mathbb{R}^{m \times n}$ have $\operatorname{rank}(\tilde{\mathbf{A}}) \leq k$. Then by the Fundamental Theorem of Linear Algebra,

$$\dim \operatorname{null}(\tilde{\mathbf{A}}) = n - \operatorname{rank}(\tilde{\mathbf{A}}) \geq n - k$$

It follows that

$$\operatorname{null}(\tilde{\mathbf{A}}) \cap \operatorname{span}\{\mathbf{v}_1, \dots, \mathbf{v}_{k+1}\}$$

is non-trivial (has a nonzero element), because otherwise there would be at least $(n-k)+(k+1) = n+1$ linearly independent vectors in $\mathbb{R}^n$, which is impossible. Therefore let $\mathbf{z}$ be some element of the intersection, and assume without loss of generality that it has unit norm: $\|\mathbf{z}\|_2 = 1$. Expand $\mathbf{z} = \alpha_1 \mathbf{v}_1 + \cdots + \alpha_{k+1} \mathbf{v}_{k+1}$, noting that

$$1 = \|\mathbf{z}\|_2^2 = \|\alpha_1 \mathbf{v}_1 + \cdots + \alpha_{k+1} \mathbf{v}_{k+1}\|_2^2 = \alpha_1^2 + \cdots + \alpha_{k+1}^2$$

by the Pythagorean theorem. Thus

$$\begin{aligned}
\|\mathbf{A} - \tilde{\mathbf{A}}\|_2 &\geq \|(\mathbf{A} - \tilde{\mathbf{A}})\mathbf{z}\|_2 && \text{by def., and } \|\mathbf{z}\|_2 = 1 \\
&= \|\mathbf{A}\mathbf{z}\|_2 && \mathbf{z} \in \operatorname{null}(\tilde{\mathbf{A}}) \\
&= \left\| \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top \mathbf{z} \right\|_2 \\
&= \left\| \sum_{i=1}^{k+1} \sigma_i \alpha_i \mathbf{u}_i \right\|_2 \\
&= \sqrt{(\sigma_1\alpha_1)^2 + \cdots + (\sigma_{k+1}\alpha_{k+1})^2} && \text{Pythagorean theorem again} \\
&\geq \sigma_{k+1}\sqrt{\alpha_1^2 + \cdots + \alpha_{k+1}^2} && \sigma_{k+1} \leq \sigma_i \text{ for } i \leq k \\
&= \|\mathbf{A} - \mathbf{A}_k\|_2 && \text{using our earlier results}
\end{aligned}$$

as was to be shown. □

A measure of the quality of the approximation is given by

$$\frac{\|\mathbf{A}_k\|_{\mathrm{F}}^2}{\|\mathbf{A}\|_{\mathrm{F}}^2} = \frac{\sigma_1^2 + \cdots + \sigma_k^2}{\sigma_1^2 + \cdots + \sigma_r^2}$$

Ideally, this ratio will be close to 1, indicating that most of the information was retained.

3.16 Pseudoinverses

Let $\mathbf{A} \in \mathbb{R}^{m \times n}$. If $m \neq n$, then $\mathbf{A}$ cannot possibly be invertible. However, there is a generalization of the inverse known as the **Moore-Penrose pseudoinverse**, denoted $\mathbf{A}^\dagger \in \mathbb{R}^{n \times m}$, which always exists and is defined uniquely by the following properties:

(i) $\mathbf{A}\mathbf{A}^\dagger\mathbf{A} = \mathbf{A}$

(ii) $\mathbf{A}^\dagger\mathbf{A}\mathbf{A}^\dagger = \mathbf{A}^\dagger$

(iii) $\mathbf{A}\mathbf{A}^\dagger$ is symmetric

(iv) $\mathbf{A}^\dagger\mathbf{A}$ is symmetric

If $\mathbf{A}$ is invertible, then $\mathbf{A}^\dagger = \mathbf{A}^{-1}$. More generally, we can compute the pseudoinverse of a matrix from its singular value decomposition: if $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, then

$$\mathbf{A}^\dagger = \mathbf{V}\mathbf{\Sigma}^\dagger\mathbf{U}^\top$$

where $\mathbf{\Sigma}^\dagger$ can be computed from $\mathbf{\Sigma}$ by taking the transpose and inverting the nonzero singular values on the diagonal. Verifying that this matrix satisfies the properties of the pseudoinverse is straightforward and left as an exercise to the reader.

3.17 Some useful matrix identities

3.17.1 Matrix-vector product as linear combination of matrix columns

Proposition 10. *Let* $\mathbf{x} \in \mathbb{R}^n$ *be a vector and* $\mathbf{A} \in \mathbb{R}^{m \times n}$ *a matrix with columns* $\mathbf{a}_1, \ldots, \mathbf{a}_n$. *Then*

$$\mathbf{A}\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{a}_i$$

This identity is extremely useful in understanding linear operators in terms of their matrices' columns. The proof is very simple (consider each element of $\mathbf{A}\mathbf{x}$ individually and expand by definitions) but it is a good exercise to convince yourself.

3.17.2 Sum of outer products as matrix-matrix product

An **outer product** is an expression of the form $\mathbf{a}\mathbf{b}^\top$, where $\mathbf{a} \in \mathbb{R}^m$ and $\mathbf{b} \in \mathbb{R}^n$. By inspection it is not hard to see that such an expression yields an $m \times n$ matrix such that

$$[\mathbf{a}\mathbf{b}^\top]_{ij} = a_i b_j$$

It is not immediately obvious, but the sum of outer products is actually equivalent to an appropriate matrix-matrix product! We formalize this statement as

Proposition 11. *Let* $\mathbf{a}_1, \ldots, \mathbf{a}_k \in \mathbb{R}^m$ *and* $\mathbf{b}_1, \ldots, \mathbf{b}_k \in \mathbb{R}^n$. *Then*

$$\sum_{\ell=1}^{k} \mathbf{a}_\ell \mathbf{b}_\ell^\top = \mathbf{A}\mathbf{B}^\top$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_k \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_k \end{bmatrix}$$

Proof. For each (i, j), we have

$$\left[\sum_{\ell=1}^{k} \mathbf{a}_\ell \mathbf{b}_\ell^\top\right]_{ij} = \sum_{\ell=1}^{k} [\mathbf{a}_\ell \mathbf{b}_\ell^\top]_{ij} = \sum_{\ell=1}^{k} [\mathbf{a}_\ell]_i [\mathbf{b}_\ell]_j = \sum_{\ell=1}^{k} A_{i\ell} B_{j\ell}$$

This last expression should be recognized as an inner product between the ith row of $\mathbf{A}$ and the jth row of $\mathbf{B}$, or equivalently the jth column of $\mathbf{B}^\top$. Hence by the definition of matrix multiplication, it is equal to $[\mathbf{A}\mathbf{B}^\top]_{ij}$. □

3.17.3 Quadratic forms

Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a symmetric matrix, and recall that the expression $\mathbf{x}^\top \mathbf{A}\mathbf{x}$ is called a quadratic form of $\mathbf{A}$. It is in some cases helpful to rewrite the quadratic form in terms of the individual elements that make up $\mathbf{A}$ and $\mathbf{x}$:

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$$

This identity is valid for any square matrix (need not be symmetric), although quadratic forms are usually only discussed in the context of symmetric matrices.