

Orthogonality and least squares

COMS 3251 Fall 2022 (Daniel Hsu)

1 Inner products and orthonormal bases

1.1 Lengths

Consider 2-vectors in the Cartesian plane as we imagine them in the real physical space. The *length* (a.k.a. *norm*) of a 2-vector $\mathbf{v} = (v_1, v_2)$, denoted by $\|\mathbf{v}\|$, has a formula provided by the Pythagorean Theorem:

$$\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2}.$$

The length $\|\mathbf{v}\|$ is the distance between the point \mathbf{v} and the origin $\mathbf{0}$, and the length $\|\mathbf{u} - \mathbf{v}\|$ is the distance between points \mathbf{u} and \mathbf{v} .

The notion of a norm generalizes to 3-vectors (displacements in three-dimensional Cartesian space) and also to n -vectors. The norm of an n -vector $\mathbf{v} = (v_1, \dots, v_n)$ is

$$\|\mathbf{v}\| = \sqrt{v_1^2 + \dots + v_n^2}.$$

Observe that $\|\mathbf{v}\| = 0$ if and only if $\mathbf{v} = \mathbf{0}$.

A *unit vector* is a vector of length 1. For example, each of the standard basis vectors $\mathbf{e}_1, \dots, \mathbf{e}_n$ is a unit vector. If $\mathbf{v} \neq \mathbf{0}$, then $\frac{1}{\|\mathbf{v}\|}\mathbf{v}$ is a unit vector.

Theorem 1 (Triangle Inequality). *For any n -vectors \mathbf{u} and \mathbf{v} ,*

$$\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|.$$

1.2 Angles and inner products

Again, consider 2-vectors in the Cartesian plane. Unit vectors correspond to points on the unit circle, which are specified by the angle between the vector and the first standard basis vector $\mathbf{e}_1 = (1, 0)$.

- If the angle between \mathbf{e}_1 and the unit vector $\mathbf{u} = (u_1, u_2)$ is $\alpha \in [0, 2\pi)$, then

$$u_1 = \cos(\alpha), \quad u_2 = \sin(\alpha).$$

- If $\mathbf{u} = (u_1, u_2) = (\cos(\alpha), \sin(\alpha))$ and $\mathbf{v} = (v_1, v_2) = (\cos(\beta), \sin(\beta))$, then the angle between \mathbf{u} and \mathbf{v} is

$$\cos(\alpha - \beta) = \cos(\alpha)\cos(\beta) + \sin(\alpha)\sin(\beta) = u_1v_1 + u_2v_2.$$

This motivates the concept of the *inner product* (a.k.a. *dot product*) between \mathbf{u} and \mathbf{v} , denoted by $\langle \mathbf{u}, \mathbf{v} \rangle$, and defined by

$$\langle \mathbf{u}, \mathbf{v} \rangle = u_1v_1 + u_2v_2.$$

(We sometimes read “ $\langle \mathbf{u}, \mathbf{v} \rangle$ ” aloud as “ \mathbf{u} dot \mathbf{v} ”.) This definition makes sense for all 2-vectors, not just the unit vectors, and its interpretation is

$$\langle \mathbf{u}, \mathbf{v} \rangle = \|\mathbf{u}\| \|\mathbf{v}\| \cos(\text{“angle between } \mathbf{u} \text{ and } \mathbf{v}\text{”}).$$

Inner products are more convenient to reason about than angles since they possess a certain property related to linearity, discussed below.

The concept of inner product generalizes to n -vectors. The inner product between $\mathbf{u} = (u_1, \dots, u_n)$ and $\mathbf{v} = (v_1, \dots, v_n)$ is defined to be

$$\langle \mathbf{u}, \mathbf{v} \rangle = u_1v_1 + \dots + u_nv_n.$$

The inner product is a real-valued, two-argument function. Moreover, it satisfies the following important properties:

IP1 (The inner product is *symmetric*.) For all vectors \mathbf{u} and \mathbf{v} ,

$$\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle.$$

IP2 (The inner product is *positive definite*.) For all vectors \mathbf{v} ,

$$\langle \mathbf{v}, \mathbf{v} \rangle \geq 0,$$

and $\langle \mathbf{v}, \mathbf{v} \rangle = 0$ if and only if $\mathbf{v} = \mathbf{0}$.

(Note that $\langle \mathbf{v}, \mathbf{v} \rangle$ gives the squared norm: $\langle \mathbf{v}, \mathbf{v} \rangle = \|\mathbf{v}\|^2$.)

IP3 (The inner product is linear in the first argument.) For all vectors \mathbf{x} , \mathbf{y} , and \mathbf{z} , and all real numbers c .

$$\langle c\mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = c\langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle.$$

IP1 and IP3 together imply that the inner product is bilinear: it is linear in each argument when the other argument value is held fixed.

The inner product also satisfies the following inequality.

Theorem 2 (Cauchy-Schwarz Inequality). *For any n -vectors \mathbf{u} and \mathbf{v} ,*

$$\langle \mathbf{u}, \mathbf{v} \rangle \leq \|\mathbf{u}\| \|\mathbf{v}\|.$$

Equality holds if and only if $\mathbf{v} = c\mathbf{u}$ for some real number c .

Example. Suppose you are given a non-zero n -vector \mathbf{x} , and you would like to find a unit vector \mathbf{v} that makes $\langle \mathbf{x}, \mathbf{v} \rangle$ as large as possible. By the Cauchy-Schwarz Inequality, the value of $\langle \mathbf{x}, \mathbf{v} \rangle$ is always at most $\|\mathbf{x}\|$, since $\|\mathbf{v}\| = 1$ for a unit vector \mathbf{v} . And we also know that the inequality holds with equality if $\mathbf{v} = c\mathbf{x}$ for some real number c . For this to hold and for \mathbf{v} to be a unit vector, it had better be that $c = 1/\|\mathbf{x}\|$. So $\mathbf{v} = \mathbf{x}/\|\mathbf{x}\|$ solves this optimization problem, and it achieves value $\langle \mathbf{x}, \mathbf{v} \rangle = \|\mathbf{x}\|$. ■

Finally, observe that if \mathbf{u}^\top is the linear functional corresponding to \mathbf{u} , then

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^\top \mathbf{v}.$$

So $\mathbf{u}^\top \mathbf{v}$ is also a commonly-used notation for inner product between n -vectors \mathbf{u} and \mathbf{v} . (Another notation is $\mathbf{u} \bullet \mathbf{v}$, to go along with the term dot product.)

1.3 Inner products for general vector spaces

Any (real) vector space \mathbb{V} may be upgraded by introducing of a real-valued, two-argument function $\langle \cdot, \cdot \rangle_{\mathbb{V}}: \mathbb{V} \times \mathbb{V} \rightarrow \mathbb{R}$ with the same properties IP1–IP3 of the inner product that we have defined for n -vectors. When we start with a vector space \mathbb{V} and then “upgrade” (or “equip”) it with a function $\langle \cdot, \cdot \rangle_{\mathbb{V}}$ satisfying IP1–IP3, we say that $(\mathbb{V}, \langle \cdot, \cdot \rangle_{\mathbb{V}})$ is a (real) inner product space.¹

The n -dimensional Cartesian space \mathbb{R}^n , equipped with the inner product we have previously defined for n -vectors (i.e., the (standard) Euclidean inner product), is called the n -dimensional Euclidean space. Henceforth, unless stated otherwise, we’ll use \mathbb{R}^n to refer to this inner product space.

¹We’ll usually just refer to \mathbb{V} itself as the inner product space, leaving implicit what the inner product is. We’ll also drop the subscript \mathbb{V} from $\langle \cdot, \cdot \rangle_{\mathbb{V}}$ when the inner product is clear from context (e.g., the standard Euclidean inner product for Euclidean space).

Example. Let $\mathbb{V} = \mathcal{C}([-1, 1], \mathbb{R})$, the space of continuous real-valued functions defined on the interval $[-1, 1]$. We equip \mathbb{V} with $\langle \cdot, \cdot \rangle_{\mathbb{V}}$, defined by

$$\langle f, g \rangle_{\mathbb{V}} = \int_{-1}^1 f(t)g(t) dt.$$

It can be verified that $\langle \cdot, \cdot \rangle_{\mathbb{V}}$ satisfies IP1–IP3. ■

Another example. Let $\mathbb{V} = \mathbb{R}^n$, but instead of considering the Euclidean inner product, we equip \mathbb{V} with $\langle \cdot, \cdot \rangle_{\mathbb{V}}$, defined by

$$\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbb{V}} = \sum_{i=1}^n \frac{1}{i^2} u_i v_i$$

for $\mathbf{u} = (u_1, \dots, u_n)$ and $\mathbf{v} = (v_1, \dots, v_n)$. Again, $\langle \cdot, \cdot \rangle_{\mathbb{V}}$ satisfies IP1–IP3. However, e.g., note that $\langle \mathbf{e}_2, \mathbf{e}_2 \rangle_{\mathbb{V}} = 1/4$ rather than 1. ■

General inner product spaces $(\mathbb{V}, \langle \cdot, \cdot \rangle_{\mathbb{V}})$ share many of the “geometric” properties we are familiar with from Euclidean space. For instance, it satisfies the Cauchy-Schwarz Inequality (Theorem 2). Moreover, we can define a notion of length based on the inner product by

$$\|\mathbf{v}\|_{\mathbb{V}} = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle_{\mathbb{V}}}.$$

Like the notion of length from Euclidean space, this notion of length $\|\cdot\|_{\mathbb{V}}$ satisfies the following properties that qualify it to be a norm:

N1 (The norm is positive definite.) For all $\mathbf{v} \in \mathbb{V}$, $\|\mathbf{v}\|_{\mathbb{V}} \geq 0$; equality holds if and only if $\mathbf{v} = \mathbf{0}_{\mathbb{V}}$.

N2 (The norm is absolutely homogeneous.) For all $\mathbf{v} \in \mathbb{V}$ and all $c \in \mathbb{R}$, $\|c\mathbf{v}\|_{\mathbb{V}} = |c| \|\mathbf{v}\|_{\mathbb{V}}$.

N3 (The norm satisfies the triangle inequality.) For all $\mathbf{u}, \mathbf{v} \in \mathbb{V}$, $\|\mathbf{u} + \mathbf{v}\|_{\mathbb{V}} \leq \|\mathbf{u}\|_{\mathbb{V}} + \|\mathbf{v}\|_{\mathbb{V}}$.

(We typically refer to $\|\mathbf{u} - \mathbf{v}\|_{\mathbb{V}}$ as the distance between \mathbf{u} and \mathbf{v} .)

Finally, much like the way linear functionals on \mathbb{R}^n are given by row vectors, each linear functional $T: \mathbb{V} \rightarrow \mathbb{R}$ on a general inner product space \mathbb{V} is uniquely specified by some vector $\mathbf{u} \in \mathbb{V}$, via $T(\mathbf{v}) = \langle \mathbf{u}, \mathbf{v} \rangle_{\mathbb{V}}$.

2 Orthogonality

2.1 Orthogonal vectors

Two vectors \mathbf{u} and \mathbf{v} from an inner product space \mathbb{V} are orthogonal (a.k.a. perpendicular) if $\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbb{V}} = 0$. Recall in the context of 2-vectors, this means that either one of the vectors is $\mathbf{0}$, or the cosine of the angle between them is 0—i.e., the angle is a right angle.

A set of vectors from an inner product space is orthogonal if every pair of distinct vectors in it is orthogonal to each other.²

Theorem 3 (Pythagorean Theorem). *Suppose $\mathbf{q}_1, \dots, \mathbf{q}_n$ are orthogonal vectors from an inner product space \mathbb{V} . Then*

$$\|\mathbf{q}_1 + \dots + \mathbf{q}_n\|_{\mathbb{V}}^2 = \|\mathbf{q}_1\|_{\mathbb{V}}^2 + \dots + \|\mathbf{q}_n\|_{\mathbb{V}}^2.$$

Proof. “Expand the square” and use orthogonality:

$$\begin{aligned} \|\mathbf{q}_1 + \dots + \mathbf{q}_n\|_{\mathbb{V}}^2 &= \langle \mathbf{q}_1 + \dots + \mathbf{q}_n, \mathbf{q}_1 + \dots + \mathbf{q}_n \rangle_{\mathbb{V}} \\ &= \sum_{i=1}^n \langle \mathbf{q}_i, \mathbf{q}_i \rangle_{\mathbb{V}} + \sum_{i=1}^n \sum_{j \neq i} \langle \mathbf{q}_i, \mathbf{q}_j \rangle_{\mathbb{V}} \overset{0}{=} \sum_{i=1}^n \|\mathbf{q}_i\|_{\mathbb{V}}^2. \quad \square \end{aligned}$$

Example. The set of 2-vectors $\{(1, 1), (2, -2)\}$ is orthogonal; the squared lengths of the vectors are 2 and 8. The sum of the vectors is $(3, -1)$, and it has squared length 10. ■

If a set (or list) of unit vectors is orthogonal, then we say it is orthonormal.

2.2 Orthogonal subspaces

If \mathbb{V} and \mathbb{W} are both subspaces of the same inner product space (e.g., \mathbb{R}^n), then we say they are orthogonal subspaces if every vector $\mathbf{v} \in \mathbb{V}$ is orthogonal to every vector $\mathbf{w} \in \mathbb{W}$.

²We say a list of vectors $(\mathbf{q}_1, \dots, \mathbf{q}_k)$ is orthogonal (or “ $\mathbf{q}_1, \dots, \mathbf{q}_k$ are orthogonal”) if they are distinct and $\{\mathbf{q}_1, \dots, \mathbf{q}_k\}$ is orthogonal.

Examples.

- Let $\mathbb{V} = \{(x, 0, 0) : x \in \mathbb{R}\}$ and $\mathbb{W} = \{(0, y, z) : (y, z) \in \mathbb{R}^2\}$ be subspaces of 3-dimensional Euclidean space. Then \mathbb{V} and \mathbb{W} are orthogonal: for any $\mathbf{v} = (v_1, v_2, v_3) \in \mathbb{V}$ and $\mathbf{w} = (w_1, w_2, w_3) \in \mathbb{W}$,

$$\langle \mathbf{v}, \mathbf{w} \rangle = v_1 w_1 + v_2 w_2 + v_3 w_3 = v_1 \cdot 0 + 0 \cdot w_2 + 0 \cdot w_3 = 0.$$

- Let $\mathbb{V} = \{(x, y, 0) : (x, y) \in \mathbb{R}^2\}$ and $\mathbb{W} = \{(0, y, z) : (y, z) \in \mathbb{R}^2\}$ be subspaces of 3-dimensional Euclidean space. Then \mathbb{V} and \mathbb{W} are not orthogonal: \mathbb{V} and \mathbb{W} both contain $\mathbf{v} = (0, 1, 0)$, and $\langle \mathbf{v}, \mathbf{v} \rangle = 1$.

Fact 1. *Orthogonal subspaces intersect only at the origin $\mathbf{0}$.*

Proof. A vector in the intersection of orthogonal subspaces must be orthogonal to itself, so the (squared) norm of the vector must be zero. \square

Proposition 1. *Let A be an $m \times n$ matrix.*

1. $\text{CS}(A^\top)$ and $\text{NS}(A)$ are orthogonal subspaces of \mathbb{R}^n .
2. $\text{CS}(A)$ and $\text{NS}(A^\top)$ are orthogonal subspaces of \mathbb{R}^m .

Proof. We just prove the first claim, as the second claim follows from the same proof after interchanging the roles of columns and rows. Consider any vector in $\text{CS}(A^\top)$; write it as $A^\top \mathbf{u}$ for some m -vector \mathbf{u} . This vector corresponds to a linear functional on \mathbb{R}^n , written as $\mathbf{u}^\top A$, so for any n -vector \mathbf{v} ,

$$\langle A^\top \mathbf{u}, \mathbf{v} \rangle = (\mathbf{u}^\top A) \mathbf{v}.$$

In particular, for any $\mathbf{v} \in \text{NS}(A)$, by associativity of matrix multiplication,³

$$(\mathbf{u}^\top A) \mathbf{v} = \mathbf{u}^\top (A \mathbf{v}) = \mathbf{u}^\top \mathbf{0} = 0. \quad (1)$$

So every vector in $\text{CS}(A^\top)$ is orthogonal to every vector in $\text{NS}(A)$. \square

³The key step $(\mathbf{u}^\top A) \mathbf{v} = \mathbf{u}^\top (A \mathbf{v})$ can be rewritten using inner products as $\langle A^\top \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{u}, A \mathbf{v} \rangle$; these are inner products in two different spaces, \mathbb{R}^n and \mathbb{R}^m . The transpose A^\top of A (changing rows of A to columns of A^\top) is the unique matrix that ensures $\langle A^\top \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{u}, A \mathbf{v} \rangle$ for all $\mathbf{u} \in \mathbb{R}^m$ and $\mathbf{v} \in \mathbb{R}^n$.

2.3 Orthogonal complements

Proposition 1 tells us that the nullspace of an $m \times n$ matrix A contains only vectors that are orthogonal to the row space $\text{CS}(A^\top)$. In fact, the nullspace contains all vectors that are orthogonal to the row space. This is, indeed, one way to interpret the definition of nullspace: $\text{NS}(A) = \{\mathbf{x} \in \mathbb{R}^n : A\mathbf{x} = \mathbf{0}\}$; it is the set of vectors orthogonal to every row of A , and hence it is the set of vectors orthogonal to every linear combination of rows of A .

For any subspace \mathbb{W} of an inner product space \mathbb{V} , define the orthogonal complement of \mathbb{W} , written \mathbb{W}^\perp (and read aloud as “ \mathbb{W} perp”), to be

$$\mathbb{W}^\perp = \{\mathbf{v} \in \mathbb{V} : \langle \mathbf{w}, \mathbf{v} \rangle_{\mathbb{V}} = 0 \text{ for all } \mathbf{w} \in \mathbb{W}\}.$$

Fact 2. *If \mathbb{W} is a subspace of an inner product space \mathbb{V} , then \mathbb{W}^\perp is a subspace of \mathbb{V} , and \mathbb{W} and \mathbb{W}^\perp are orthogonal subspaces.*

Proof. The proof that \mathbb{W}^\perp is a subspace of \mathbb{V} is completely analogous to the proof that $\text{NS}(A)$ is a subspace for any matrix A . The fact that \mathbb{W} and \mathbb{W}^\perp are orthogonal follows by definition. \square

In this notation, we have $\text{NS}(A) = \text{CS}(A^\top)^\perp$: the nullspace of A is the orthogonal complement of the row space of A . In fact, it is also the case that the row space is the orthogonal complement of the nullspace.

Theorem 4. *Let A be an $m \times n$ matrix.*

1. $\text{NS}(A) = \text{CS}(A^\top)^\perp$ and $\text{CS}(A^\top) = \text{NS}(A)^\perp$.
2. $\text{NS}(A^\top) = \text{CS}(A)^\perp$ and $\text{CS}(A) = \text{NS}(A^\top)^\perp$.

Proof. We already saw that $\text{NS}(A) = \text{CS}(A^\top)^\perp$, essentially by definition. We now prove that $\text{CS}(A^\top) = \text{NS}(A)^\perp$. Suppose for sake of contradiction that there exists a vector $\mathbf{v} \in \mathbb{R}^n$ that is orthogonal to every vector in the nullspace of A , and yet $\mathbf{v} \notin \text{CS}(A^\top)$. Consider the matrix B that is the same as A with an additional row \mathbf{v}^\top . Since $\mathbf{v} \notin \text{CS}(A^\top)$, the Growth Theorem implies that the dimension of the row space of B is one more than the dimension of the row space of A : $\text{rank}(B) = \text{rank}(A) + 1$. On the other hand, the nullspace of B is the same as the nullspace of A , since \mathbf{v} is orthogonal to every vector in $\text{NS}(A)$. Using the Dimension Theorem with B tells us

$$\text{rank}(B) + \dim(\text{NS}(B)) = \text{rank}(A) + 1 + \dim(\text{NS}(A)) = n,$$

but using it with A tells us $\text{rank}(A) + \dim(\text{NS}(A)) = n$. This is a contradiction, so we conclude no such vector \mathbf{v} exists. Hence $\text{CS}(A^\top) = \text{NS}(A)^\perp$.

Switching the roles of rows and columns proves the second claim. \square

3 Orthonormal bases and orthoprojectors

3.1 Orthonormal bases

Recall that a basis for a vector space \mathbb{V} is a minimal collection of vectors by which you can construct all of \mathbb{V} simply via linear combination. If \mathbb{V} is, in fact, an inner product space, then bases that are orthonormal (i.e., composed of orthonormal vectors) are especially convenient. We use the term orthonormal basis (ONB) for a (ordered) basis that is orthonormal.

Examples. The standard ordered basis $(\mathbf{e}_1, \dots, \mathbf{e}_n)$ is an ONB for \mathbb{R}^n . For $n = 2$, this is

$$\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right).$$

Pick any $\theta \in [0, 2\pi)$. Then

$$\left(\begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix}, \begin{bmatrix} -\sin(\theta) \\ \cos(\theta) \end{bmatrix} \right)$$

is also an ONB for \mathbb{R}^2 . Each vector is a unit vector, since $\cos(\theta)^2 + \sin(\theta)^2 = 1$ for any θ . And the vectors are clearly orthogonal. \blacksquare

Very important example. Consider the vector space $\mathbb{V} = \text{C}_{\text{periodic}}([0, 2\pi], \mathbb{R})$ of continuous, real-valued functions on $[0, 2\pi]$ that are periodic with period 2π , equipped with the inner product

$$\langle f, g \rangle_{\mathbb{V}} = \frac{1}{2\pi} \int_0^{2\pi} f(t)g(t) dt.$$

The following set behaves much like an ONB for \mathbb{V} :

$$\{1\} \cup \left\{ \sqrt{2} \cos(kt) : k \in \mathbb{N} \right\} \cup \left\{ \sqrt{2} \sin(kt) : k \in \mathbb{N} \right\}.$$

A bit of calculus verifies that each function has norm 1, and also that every distinct pair is orthogonal. The reason it is technically not a basis is because to express some functions in \mathbb{V} , we may need to linearly combine infinitely-many basis vectors. Such representations of periodic functions are called Fourier series. Here are two examples:

$$t(t - \pi)(t - 2\pi) = 6\sqrt{2} \sum_{k=1}^{\infty} \frac{1}{k^3} \sqrt{2} \sin(kt);$$

$$\min\{t/\pi, 2 - t/\pi\} = \frac{1}{2} - \frac{2\sqrt{2}}{\pi^2} \sum_{\text{odd } k=1}^{\infty} \frac{1}{k^2} \sqrt{2} \cos(kt).$$

(Try plotting finite prefixes of these series.) These representations are obtained using the method described in the theorem below, which converts between the time domain (values $f(t)$ for every “time” t) and the frequency domain (coefficients of sines and cosines in its Fourier series). ■

The following theorem shows how to obtain the coordinate representation of a vector from an inner product space with respect to a basis of non-zero orthogonal vectors.

Theorem 5. Let $\{\mathbf{q}_1, \dots, \mathbf{q}_k\}$ be an orthogonal set of k non-zero vectors from an inner product space \mathbb{V} . If $\mathbf{x} = c_1\mathbf{q}_1 + \dots + c_k\mathbf{q}_k$ for some scalars c_1, \dots, c_k , then

$$c_i = \frac{\langle \mathbf{x}, \mathbf{q}_i \rangle_{\mathbb{V}}}{\|\mathbf{q}_i\|_{\mathbb{V}}^2} \quad \text{for all } i \in \{1, \dots, k\}.$$

Proof. By linearity of the inner product and orthogonality of $\{\mathbf{q}_1, \dots, \mathbf{q}_k\}$,

$$\langle \mathbf{x}, \mathbf{q}_i \rangle_{\mathbb{V}} = c_1 \langle \mathbf{q}_1, \mathbf{q}_i \rangle_{\mathbb{V}} + \dots + c_k \langle \mathbf{q}_k, \mathbf{q}_i \rangle_{\mathbb{V}} = c_i \langle \mathbf{q}_i, \mathbf{q}_i \rangle_{\mathbb{V}} = c_i \|\mathbf{q}_i\|_{\mathbb{V}}^2$$

for each $i \in \{1, \dots, k\}$. Solve for each c_i proves the claim. □

The following corollary specializes to the case of an ONB.

Corollary 1. Let $\{\mathbf{q}_1, \dots, \mathbf{q}_n\}$ be an ONB for an n -dimensional inner product space \mathbb{V} . For every $\mathbf{x} \in \mathbb{V}$,

$$\mathbf{x} = \langle \mathbf{x}, \mathbf{q}_1 \rangle_{\mathbb{V}} \mathbf{q}_1 + \dots + \langle \mathbf{x}, \mathbf{q}_n \rangle_{\mathbb{V}} \mathbf{q}_n$$

and $\|\mathbf{x}\|_{\mathbb{V}}^2 = \langle \mathbf{x}, \mathbf{q}_1 \rangle_{\mathbb{V}}^2 + \dots + \langle \mathbf{x}, \mathbf{q}_n \rangle_{\mathbb{V}}^2.$

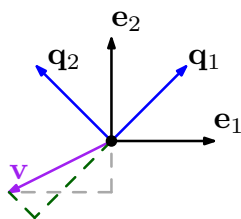


Figure 1: The dashed lines suggest how to compute the length of the 2-vector \mathbf{v} two different ways: the gray lines use the ONB $(\mathbf{e}_1, \mathbf{e}_2)$, the green lines use the ONB $(\mathbf{q}_1, \mathbf{q}_2)$.

Proof. The first identity is immediate from Theorem 5 and the assumption that the \mathbf{q}_i 's are unit vectors. Let $c_i = \langle \mathbf{q}_i, \mathbf{x} \rangle_{\mathbb{V}}$ for each $i \in \{1, \dots, n\}$. By the Pythagorean Theorem (Theorem 3) and absolute homogeneity,

$$\begin{aligned} \|\mathbf{x}\|_{\mathbb{V}}^2 &= \|c_1 \mathbf{q}_1 + \dots + c_n \mathbf{q}_n\|_{\mathbb{V}}^2 = \|c_1 \mathbf{q}_1\|_{\mathbb{V}}^2 + \dots + \|c_n \mathbf{q}_n\|_{\mathbb{V}}^2 \\ &= c_1^2 \|\mathbf{q}_1\|_{\mathbb{V}}^2 + \dots + c_n^2 \|\mathbf{q}_n\|_{\mathbb{V}}^2. \quad \square \end{aligned}$$

(The second identity in Corollary 1 is known as Parseval's identity.)

Example. Let $\mathbb{V} = \mathbb{R}^2$, and for some $\theta \in [0, 2\pi)$, consider the ordered basis $\mathcal{Q} = (\mathbf{q}_1, \mathbf{q}_2)$, where $\mathbf{q}_1 = (\cos(\theta), \sin(\theta))$ and $\mathbf{q}_2 = (-\sin(\theta), \cos(\theta))$. The vector $\mathbf{v} = (3, 4)$ has squared norm $3^2 + 4^2 = 25$; it can also be computed as

$$\begin{aligned} \langle \mathbf{q}_1, \mathbf{v} \rangle^2 + \langle \mathbf{q}_2, \mathbf{v} \rangle^2 &= (3 \cos(\theta) + 4 \sin(\theta))^2 + (-3 \sin(\theta) + 4 \cos(\theta))^2 \\ &= (9 + 16)(\sin^2(\theta) + \cos^2(\theta)) = 25. \end{aligned}$$

See Figure 1 for another example. ■

Corollary 1 shows that, for orthonormal bases, getting the coordinate representation of a vector is conceptually simple:

$$[\mathbf{x}]_{\mathcal{Q}} = \begin{bmatrix} \langle \mathbf{q}_1, \mathbf{x} \rangle_{\mathbb{V}} \\ \vdots \\ \langle \mathbf{q}_n, \mathbf{x} \rangle_{\mathbb{V}} \end{bmatrix},$$

where $\mathcal{Q} = (\mathbf{q}_1, \dots, \mathbf{q}_n)$ is the (ordered) ONB for \mathbb{V} . The coordinates also provide another way to compute the squared norm:

$$\|\mathbf{x}\|_{\mathbb{V}}^2 = \|[x]_{\mathcal{Q}}\|^2;$$

the right-hand side norm is the standard Euclidean norm for n -vectors.

If, in the same context as above, $\mathbb{V} = \mathbb{R}^n$ and $Q = [\mathbf{q}_1, \dots, \mathbf{q}_n]$ is the $n \times n$ matrix with the basis vectors as columns, then $[\mathbf{x}]_{\mathcal{Q}} = Q^T \mathbf{x}$.⁴ It is clear that $[\mathbf{q}_i]_{\mathcal{Q}} = \mathbf{e}_i$ for each $i \in \{1, \dots, n\}$, and therefore $Q^T Q = I$. Moreover, for any $\mathbf{x} \in \mathbb{R}^n$, we have $Q Q^T \mathbf{x} = Q [\mathbf{x}]_{\mathcal{Q}} = \mathbf{x}$, so $Q Q^T = I$ as well. This shows that Q is invertible, and its inverse is $Q^{-1} = Q^T$. A square matrix with orthonormal columns is called an *orthogonal matrix*.⁵

Below is a related corollary of Theorem 5 for general inner product spaces.

Corollary 2. *If $\{\mathbf{q}_1, \dots, \mathbf{q}_k\}$ is an orthogonal set of k non-zero vectors from an inner product space \mathbb{V} , then it is linearly independent.*

Proof. Apply Theorem 5 with $\mathbf{x} = \mathbf{0}$ to deduce that any linear combination of distinct vectors from $\{\mathbf{q}_1, \dots, \mathbf{q}_k\}$ must be the all-zeros combination. \square

Corollary 2, with the Subspace Dimension Theorem, implies that every orthonormal subset of an inner product space of dimension n has cardinality at most n .

3.2 Gram-Schmidt orthogonalization

The following algorithm takes as input linearly independent vectors from an inner product space, and returns an orthogonal set of non-zero vectors that has the same span. To get an ONB for the span, divide each vector in the output by its norm.

Algorithm 1 Gram-Schmidt orthogonalization

Input: Linearly independent vectors $\mathbf{b}_1, \dots, \mathbf{b}_d$ from inner product space \mathbb{V} .

1: **for** $k = 1, \dots, d$ **do**

2: Let $\mathbf{q}_k = \mathbf{b}_k - \sum_{j=1}^{k-1} \frac{\langle \mathbf{b}_k, \mathbf{q}_j \rangle_{\mathbb{V}}}{\|\mathbf{q}_j\|_{\mathbb{V}}^2} \mathbf{q}_j$.

3: **end for**

4: **return** $\{\mathbf{q}_1, \dots, \mathbf{q}_d\}$.

⁴In the special case where \mathcal{Q} is the standard ordered basis, we have $Q = I$. So the n -vector itself is its own coordinate representation with respect to the standard basis.

⁵That was not a typo. An $n \times n$ matrix with n orthonormal columns is called an “orthogonal matrix”, not “orthonormal matrix”. Confusing ...

The summation in Line 2 of Algorithm 1 can be recognized as the “part” of \mathbf{b}_k that is in the span of $\{\mathbf{q}_1, \dots, \mathbf{q}_{k-1}\}$, so \mathbf{q}_k is set to the remaining “part” of \mathbf{b}_k . Precisely what these “parts” are will be explained in the context of orthogonal projections later.

Example. Consider the execution of Algorithm 1 on the following vectors:

$$[\mathbf{b}_1 \ \mathbf{b}_2 \ \mathbf{b}_3] = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 0 & 2 \\ 0 & 1 & 1 \end{bmatrix}.$$

- Iteration $k = 1$:

$$\mathbf{q}_1 = \mathbf{b}_1 = (1, 1, 0).$$

(The sum from $j = 1$ to 0 is the empty sum.)

- Iteration $k = 2$:

$$\mathbf{q}_2 = \mathbf{b}_2 - \frac{\langle \mathbf{b}_2, \mathbf{q}_1 \rangle}{\|\mathbf{q}_1\|^2} \mathbf{q}_1 = (2, 0, 1) - \frac{2}{2} (1, 1, 0) = (1, -1, 1).$$

- Iteration $k = 3$:

$$\begin{aligned} \mathbf{q}_3 &= \mathbf{b}_3 - \frac{\langle \mathbf{b}_3, \mathbf{q}_1 \rangle}{\|\mathbf{q}_1\|^2} \mathbf{q}_1 - \frac{\langle \mathbf{b}_3, \mathbf{q}_2 \rangle}{\|\mathbf{q}_2\|^2} \mathbf{q}_2 \\ &= (2, 2, 1) - \frac{4}{2} (1, 1, 0) - \frac{1}{3} (1, -1, 1) \\ &= \left(-\frac{1}{3}, \frac{1}{3}, \frac{2}{3} \right). \quad \blacksquare \end{aligned}$$

Theorem 6. *The execution of Algorithm 1 on d linearly independent vectors $\mathcal{B} = \{\mathbf{b}_1, \dots, \mathbf{b}_d\}$ from an inner product space \mathbb{V} returns an orthogonal set $\mathcal{Q} = \{\mathbf{q}_1, \dots, \mathbf{q}_d\}$ of d non-zero vectors with $\text{span}(\mathcal{Q}) = \text{span}(\mathcal{B})$.*

Proof. The proof is by induction on d . The base case $d = 0$ is trivial since $\mathcal{B} = \mathcal{Q} = \emptyset$. So, for some $d \geq 1$, assume as the inductive hypothesis that $\mathcal{Q}^- = \{\mathbf{q}_1, \dots, \mathbf{q}_{d-1}\}$ is an orthogonal set of $d - 1$ non-zero vectors with $\text{span}(\mathcal{Q}^-) = \text{span}(\{\mathbf{b}_1, \dots, \mathbf{b}_{d-1}\})$. We need to show (i) \mathbf{q}_d is non-zero, (ii) $\mathcal{Q}^- \cup \{\mathbf{q}_d\}$ is orthogonal, and (iii) $\text{span}(\mathcal{Q}^- \cup \{\mathbf{q}_d\}) = \text{span}(\mathcal{B})$.

To prove (i), we assume for sake of contradiction that $\mathbf{q}_d = \mathbf{0}$. Then Line 2 in Algorithm 1 shows that $\mathbf{b}_d \in \text{span}(\mathcal{Q}^-)$, and we know $\text{span}(\mathcal{Q}^-) = \text{span}(\{\mathbf{b}_1, \dots, \mathbf{b}_{d-1}\})$ by the inductive hypothesis. This implies that the set $\{\mathbf{b}_1, \dots, \mathbf{b}_d\}$ is linearly dependent, a contradiction. So we conclude $\mathbf{q}_d \neq \mathbf{0}$.

To prove (ii), it suffices to show that $\langle \mathbf{q}_d, \mathbf{q}_k \rangle_{\mathbb{V}} = 0$ for each $k \in \{1, \dots, d-1\}$. For each such k , using linearity of the inner product and the orthogonality of \mathcal{Q}^- from the inductive hypothesis, we have

$$\begin{aligned} \langle \mathbf{q}_d, \mathbf{q}_k \rangle_{\mathbb{V}} &= \langle \mathbf{b}_d, \mathbf{q}_k \rangle_{\mathbb{V}} - \sum_{j=1}^{d-1} \frac{\langle \mathbf{b}_d, \mathbf{q}_j \rangle_{\mathbb{V}}}{\|\mathbf{q}_j\|_{\mathbb{V}}^2} \langle \mathbf{q}_j, \mathbf{q}_k \rangle_{\mathbb{V}} \\ &= \langle \mathbf{b}_d, \mathbf{q}_k \rangle_{\mathbb{V}} - \frac{\langle \mathbf{b}_d, \mathbf{q}_k \rangle_{\mathbb{V}}}{\|\mathbf{q}_k\|_{\mathbb{V}}^2} \langle \mathbf{q}_k, \mathbf{q}_k \rangle_{\mathbb{V}} = \langle \mathbf{b}_d, \mathbf{q}_k \rangle_{\mathbb{V}} - \langle \mathbf{b}_d, \mathbf{q}_k \rangle_{\mathbb{V}} = 0. \end{aligned}$$

Finally, to prove (iii), note that $\text{span}(\mathcal{Q}^- \cup \{\mathbf{q}_d\}) \subseteq \text{span}(\mathcal{B})$ follows from the inductive hypothesis that $\text{span}(\mathcal{Q}^-) = \text{span}(\{\mathbf{b}_1, \dots, \mathbf{b}_{d-1}\})$ and the fact $\mathbf{q}_d \in \text{span}(\mathcal{Q}^- \cup \mathcal{B})$. We have shown, in (i) and (ii), that $\mathcal{Q}^- \cup \{\mathbf{q}_d\}$ is an orthogonal set of non-zero vectors, and hence it is linearly independent by Corollary 2. This implies $\dim(\text{span}(\mathcal{Q}^- \cup \{\mathbf{q}_d\})) = d = \dim(\text{span}(\mathcal{B}))$, so the Subspace Dimension Theorem implies that $\text{span}(\mathcal{Q}^- \cup \{\mathbf{q}_d\}) = \text{span}(\mathcal{B})$.

This completes the inductive step, and hence the claim follows by the principle of mathematical induction. \square

Corollary 3. *If \mathbb{V} is an n -dimensional inner product space, then \mathbb{V} has an orthonormal basis.*

Proof. Apply Algorithm 1 to a basis for \mathbb{V} (which has n vectors). By Theorem 6, the output $\{\mathbf{q}_1, \dots, \mathbf{q}_n\}$ is an orthogonal set of non-zero vectors that also spans \mathbb{V} . Let $\mathbf{u}_i = \mathbf{q}_i / \|\mathbf{q}_i\|_{\mathbb{V}}$ for each $i \in \{1, \dots, n\}$, so $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ is an ONB for \mathbb{V} . \square

There is an analogue of the Basis Completion Theorem for ONB's. If a set of vectors \mathcal{W} from an inner product space \mathbb{V} is orthonormal, then it can be “completed” to become an ONB. This is done as follows:

1. Use the Basis Completion Theorem with \mathcal{W} to obtain a basis \mathcal{B} for \mathbb{V} that includes \mathcal{W} . (Note that \mathcal{B} might not be orthogonal.)
2. Apply Gram-Schmidt orthogonalization (Algorithm 1) on this basis \mathcal{B} , starting with the vectors from \mathcal{W} .

Since the vectors in \mathcal{W} are already orthogonal, they will be taken as-is as part of the output. This proves the following theorem.

Theorem 7 (ONB Completion Theorem). *Let \mathcal{W} be an orthonormal set of k vectors from an n -dimensional inner product space \mathbb{V} . There exists a subset \mathcal{F} of $n - k$ vectors such that $\mathcal{W} \cup \mathcal{F}$ is an ONB for \mathbb{V} .*

3.3 Orthogonal projections

We say that a vector space \mathbb{V} is the direct sum of its subspaces \mathbb{W}_1 and \mathbb{W}_2 , written $\mathbb{V} = \mathbb{W}_1 \oplus \mathbb{W}_2$, if for every $\mathbf{x} \in \mathbb{V}$, there exists unique choices of $\mathbf{y} \in \mathbb{W}_1$ and $\mathbf{z} \in \mathbb{W}_2$ such that $\mathbf{x} = \mathbf{y} + \mathbf{z}$.

Theorem 8 (Direct Sum Theorem). *Let \mathbb{V} be a finite dimensional inner product space, and let \mathbb{W} be a subspace of \mathbb{V} . Then $\mathbb{V} = \mathbb{W} \oplus \mathbb{W}^\perp$.*

Proof. Let $\mathcal{W} = \{\mathbf{w}_1, \dots, \mathbf{w}_k\}$ be an ONB for \mathbb{W} . By the ONB Completion Theorem (Theorem 7), there exists a subset $\mathcal{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_\ell\}$ such that $\mathcal{W} \cup \mathcal{V}$ is an ONB for \mathbb{V} (where $k + \ell = \dim(\mathbb{V})$). For any $\mathbf{x} \in \mathbb{V}$, by Corollary 1,

$$\mathbf{x} = \underbrace{\langle \mathbf{x}, \mathbf{w}_1 \rangle \mathbf{w}_1 + \dots + \langle \mathbf{x}, \mathbf{w}_k \rangle \mathbf{w}_k}_{\mathbf{y}} + \underbrace{\langle \mathbf{x}, \mathbf{v}_1 \rangle \mathbf{v}_1 + \dots + \langle \mathbf{x}, \mathbf{v}_\ell \rangle \mathbf{v}_\ell}_{\mathbf{z}}. \quad (2)$$

It is clear that $\mathbf{y} \in \mathbb{W}$; moreover, $\mathbf{z} \in \mathbb{W}^\perp$ since every \mathbf{v}_i is orthogonal to every vector in \mathbb{W} . So the existence of the claimed \mathbf{y} and \mathbf{z} with $\mathbf{x} = \mathbf{y} + \mathbf{z}$ is proven. To show uniqueness, suppose $\mathbf{x} = \mathbf{y}' + \mathbf{z}'$ for some $\mathbf{y}' \in \mathbb{W}$ and $\mathbf{z}' \in \mathbb{W}^\perp$. Then $\mathbf{y} - \mathbf{y}' = \mathbf{z}' - \mathbf{z}$. But $\mathbf{y} - \mathbf{y}' \in \mathbb{W}$ and $\mathbf{z}' - \mathbf{z} \in \mathbb{W}^\perp$, since \mathbb{W} and \mathbb{W}^\perp are both subspaces of \mathbb{R}^n . Since $\mathbb{W} \cap \mathbb{W}^\perp = \{\mathbf{0}\}$, it follows that $\mathbf{y} = \mathbf{y}'$ and $\mathbf{z} = \mathbf{z}'$. \square

For any subspace \mathbb{W} of a finite-dimensional inner product space \mathbb{V} , Theorem 8 uniquely decomposes every $\mathbf{x} \in \mathbb{V}$ into the sum of a “part” \mathbf{y} that lives in \mathbb{W} and an orthogonal “part” $\mathbf{z} = \mathbf{x} - \mathbf{y}$ that lives in \mathbb{W}^\perp . The proof shows how to extract these “parts”: obtain an ONB for \mathbb{W} , compute \mathbf{y} as shown in (2), and set $\mathbf{z} = \mathbf{x} - \mathbf{y}$. We say \mathbf{y} is the orthogonal projection of \mathbf{x} to \mathbb{W} .

Example. Let $\mathbb{W} = \text{span}(\{\mathbf{e}_1, \mathbf{e}_2\})$, a two-dimensional subspace of \mathbb{R}^3 . The orthogonal projection of $\mathbf{x} = (1, 2, 3) = \mathbf{e}_1 + 2\mathbf{e}_2 + 3\mathbf{e}_3$ to \mathbb{W} is $\mathbf{y} = (1, 2, 0) = \mathbf{e}_1 + 2\mathbf{e}_2$. Notice that $\mathbf{x} - \mathbf{y} = (0, 0, 3) = 3\mathbf{e}_3 \in \mathbb{W}^\perp$, and

$$\|\mathbf{x}\|^2 = 1^2 + 2^2 + 3^2 = \|\mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2. \quad \blacksquare$$

Let's specialize to $\mathbb{V} = \mathbb{R}^n$. The linear operator P that sends an arbitrary $\mathbf{x} \in \mathbb{R}^n$ to the unique $\mathbf{y} = P\mathbf{x} \in \mathbb{W}$ such that $\mathbf{x} - \mathbf{y} \in \mathbb{W}^\perp$ is called the orthogonal projection operator (a.k.a. orthogonal projector, orthoprojector) for \mathbb{W} . Note that $I - P$ is the orthoprojector for \mathbb{W}^\perp , by symmetry. Both P and $I - P$ are projection operators, in the sense that each is idempotent: $P^2 = P$ and $(I - P)^2 = I - P$.

It is clear an orthoprojector P is a linear operator on \mathbb{R}^n : if $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ is an ONB for \mathbb{W} (so $r = \dim(\mathbb{W})$), then

$$\begin{aligned} P\mathbf{x} &= \langle \mathbf{x}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \dots + \langle \mathbf{x}, \mathbf{u}_r \rangle \mathbf{u}_r \\ &= \begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{u}_1 & \dots & \mathbf{u}_r \\ \downarrow & & \downarrow \end{bmatrix} \begin{bmatrix} \langle \mathbf{x}, \mathbf{u}_1 \rangle \\ \vdots \\ \langle \mathbf{x}, \mathbf{u}_r \rangle \end{bmatrix} = \underbrace{\begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{u}_1 & \dots & \mathbf{u}_r \\ \downarrow & & \downarrow \end{bmatrix}}_U \underbrace{\begin{bmatrix} \leftarrow \mathbf{u}_1^\top \rightarrow \\ \vdots \\ \leftarrow \mathbf{u}_r^\top \rightarrow \end{bmatrix}}_{U^\top} \begin{bmatrix} \uparrow \\ \mathbf{x} \\ \downarrow \end{bmatrix}. \end{aligned}$$

So $P = UU^\top$, where U is the $n \times r$ matrix whose columns form an ONB for the subspace \mathbb{W} . Another way to write UU^\top is as a sum of r outer products:

$$P = UU^\top = \mathbf{u}_1\mathbf{u}_1^\top + \dots + \mathbf{u}_r\mathbf{u}_r^\top.$$

If $r = 1$, then we can recognize $P = \mathbf{u}_1\mathbf{u}_1^\top$ as a special case of an elementary projection operator to a line along a hyperplane. In this special case, the line $\text{CS}(\mathbf{u}_1) = \{c\mathbf{u}_1 : c \in \mathbb{R}\}$ and hyperplane $\text{NS}(\mathbf{u}_1^\top) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{u}_1^\top\mathbf{x} = 0\}$ are orthogonal complements of each other.

The orthoprojector for a subspace \mathbb{W} is not specific to any particular ONB. So if $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ and $\{\mathbf{w}_1, \dots, \mathbf{w}_r\}$ are both ONB's for \mathbb{W} , then

$$\mathbf{u}_1\mathbf{u}_1^\top + \dots + \mathbf{u}_r\mathbf{u}_r^\top = \mathbf{w}_1\mathbf{w}_1^\top + \dots + \mathbf{w}_r\mathbf{w}_r^\top.$$

We conclude with a very important theorem.

Theorem 9. *Let A be an $m \times n$ matrix. For every $\mathbf{b} \in \text{CS}(A)$, there exists a unique $\mathbf{y} \in \text{CS}(A^\top)$ such that $\mathbf{b} = A\mathbf{y}$.*

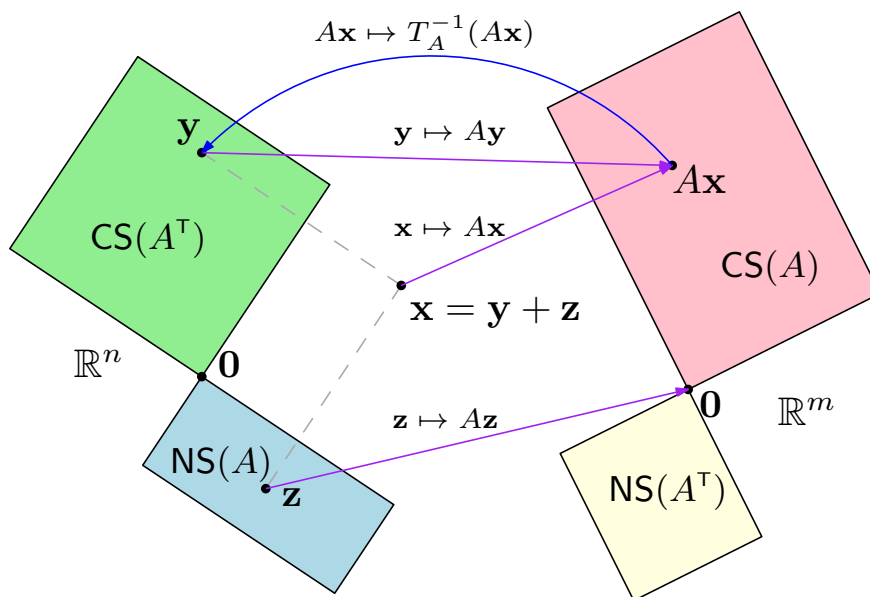


Figure 2: Schematic diagram of the fundamental subspaces of an $m \times n$ matrix A and its action on $\mathbf{x} \in \mathbb{R}^n$. Here, $T_A: \text{CS}(A^\top) \rightarrow \text{CS}(A)$ is the bijection between $\text{CS}(A^\top)$ and $\text{CS}(A)$, and $T_A^{-1}: \text{CS}(A) \rightarrow \text{CS}(A^\top)$ is its inverse.

Proof. Fix any $\mathbf{b} \in \text{CS}(A)$, so there exists $\mathbf{x} \in \mathbb{R}^n$ such that $\mathbf{b} = A\mathbf{x}$. Let \mathbf{y} be the orthogonal projection of \mathbf{x} to $\text{CS}(A^\top)$, so $\mathbf{z} = \mathbf{x} - \mathbf{y} \in \text{NS}(A)$. By linearity, $A\mathbf{y} = A(\mathbf{x} - \mathbf{z}) = A\mathbf{x} - A\mathbf{z} = A\mathbf{x}$. This proves the existence of the vector $\mathbf{y} \in \text{CS}(A^\top)$ with $\mathbf{b} = A\mathbf{y}$.

Now we prove the uniqueness of \mathbf{y} . Consider any $\mathbf{x} \in \text{CS}(A^\top)$ such that $A\mathbf{x} = \mathbf{b}$. Then $A(\mathbf{x} - \mathbf{y}) = \mathbf{0}$, so $\mathbf{x} - \mathbf{y} \in \text{NS}(A)$. On the other hand, $\mathbf{x} - \mathbf{y} \in \text{CS}(A^\top)$. But $\text{CS}(A^\top) \cap \text{NS}(A) = \{\mathbf{0}\}$ since $\text{CS}(A^\top) = \text{NS}(A)^\perp$, so it must be that $\mathbf{x} = \mathbf{y}$. \square

Theorem 9 implies that the linear transformation $T_A: \text{CS}(A^\top) \rightarrow \text{CS}(A)$ given by $T_A(\mathbf{x}) = A\mathbf{x}$ is bijjective, i.e., one-to-one and onto. See Figure 2.

4 Least squares approximation

In the least squares approximation problem, one is given an $n \times p$ matrix A and an n -vector \mathbf{b} , and the goal is to find a p -vector \mathbf{x} that makes $\|A\mathbf{x} - \mathbf{b}\|^2$ as small as possible.

In statistics, this problem is called *least squares linear regression*, which is motivated as follows. The matrix A is a coefficient matrix for a system of n linear equations in p variables $\mathbf{x} = (x_1, \dots, x_p)$, and the vector \mathbf{b} is the vector of right-hand side values. We would like to find a solution to the system $A\mathbf{x} = \mathbf{b}$ —i.e., a setting of the p variables (x_1, \dots, x_p) that satisfies all n equations—but in the case that the system is inconsistent, we would like to assign values to the p variables to make all of the equations as “close” to being satisfied as possible. The quality of an assignment is judged by the sum of the squared *residuals* for the n equations. If \mathbf{a}_i^\top is the i th row of A and b_i is the i th component of \mathbf{b} , then the i th residual of our proposed assignment is $b_i - \mathbf{a}_i^\top \mathbf{x}$. So the sum of squared residuals is

$$(b_1 - \mathbf{a}_1^\top \mathbf{x})^2 + \cdots + (b_n - \mathbf{a}_n^\top \mathbf{x})^2 = \|\mathbf{b} - A\mathbf{x}\|^2.$$

Here is one approach to solving the least squares approximation problem.

1. Compute the orthogonal projection of \mathbf{b} to $\text{CS}(A)$.

We’ve seen the steps for getting the orthoprojector P for $\text{CS}(A)$ in Section 3.3. The key step involves obtaining an ONB for $\text{CS}(A)$ via, say, Gram-Schmidt orthogonalization (Algorithm 1).

Let $\mathbf{b}_0 = P\mathbf{b}$ denote the application of P to \mathbf{b} , i.e., the orthogonal projection of \mathbf{b} to $\text{CS}(A)$.

2. Since $\mathbf{b}_0 \in \text{CS}(A)$, we simply need to solve the system of linear equations $A\mathbf{x} = \mathbf{b}_0$, which is guaranteed to have a solution. This can be done using Elimination.

Why does this work? We need to show that among all vectors in $\text{CS}(A)$, the orthogonal projection of \mathbf{b} to $\text{CS}(A)$ is the one closest to \mathbf{b} . This is the content of the next theorem.

Theorem 10. *Let \mathbb{W} be a subspace of \mathbb{R}^n , and let \mathbf{b} denote any n -vector. If $P\mathbf{b}$ is the orthogonal projection of \mathbf{b} to \mathbb{W} , then for any $\mathbf{w} \in \mathbb{W}$,*

$$\|\mathbf{b} - \mathbf{w}\|^2 = \|P\mathbf{b} - \mathbf{w}\|^2 + \|\mathbf{b} - P\mathbf{b}\|^2 \geq \|\mathbf{b} - P\mathbf{b}\|^2,$$

where the inequality holds with equality if and only if $\mathbf{w} = P\mathbf{b}$.

Proof. Write $\mathbf{b} - \mathbf{w} = P(\mathbf{b} - \mathbf{w}) + (I - P)(\mathbf{b} - \mathbf{w})$. Note that $P(\mathbf{b} - \mathbf{w}) \in \mathbb{W}$ and $(I - P)(\mathbf{b} - \mathbf{w}) \in \mathbb{W}^\perp$, so by the Pythagorean Theorem (Theorem 3),

$$\|\mathbf{b} - \mathbf{w}\|^2 = \|P(\mathbf{b} - \mathbf{w})\|^2 + \|(I - P)(\mathbf{b} - \mathbf{w})\|^2. \quad (3)$$

Since $\mathbf{w} \in \mathbb{W}$, it follows that $P\mathbf{w} = \mathbf{w}$ and $(I - P)\mathbf{w} = \mathbf{0}$. Therefore $P(\mathbf{b} - \mathbf{w}) = P\mathbf{b} - \mathbf{w}$ and $(I - P)(\mathbf{b} - \mathbf{w}) = \mathbf{b} - P\mathbf{b}$. Plugging back into (3), we get $\|\mathbf{b} - \mathbf{w}\|^2 = \|P\mathbf{b} - \mathbf{w}\|^2 + \|\mathbf{b} - P\mathbf{b}\|^2$, which is always at least $\|\mathbf{b} - P\mathbf{b}\|^2$ since the norm is non-negative. The fact that equality holds if and only if $\mathbf{w} = P\mathbf{b}$ follows by the positive definiteness of the norm. \square

The two-stage procedure we described for solving the least squares approximation problem is a bit roundabout, especially if the ONB for $\text{CS}(A)$ is not needed for anything else. A more direct approach is motivated as follows.

- We are seeking the unique vector $\mathbf{b}_0 \in \text{CS}(A)$ such that $\mathbf{b} - \mathbf{b}_0$ is orthogonal to every vector in $\text{CS}(A)$. (This is what it means for \mathbf{b}_0 to be the orthogonal projection of \mathbf{b} to $\text{CS}(A)$, as we have discussed above.) Since $\mathbf{b}_0 \in \text{CS}(A)$, we know there is a p -vector \mathbf{x} such that $\mathbf{b}_0 = A\mathbf{x}$.
- Every vector in $\text{CS}(A)$ is a linear combination of columns of A . Therefore, for $A\mathbf{x}$ to be the orthogonal projection of \mathbf{b} to $\text{CS}(A)$, it is equivalent to ensure that $\mathbf{b} - A\mathbf{x}$ is orthogonal to every column of $\text{CS}(A)$. This condition can be expressed using matrix-vector multiplication:

$$A^\top(\mathbf{b} - A\mathbf{x}) = \mathbf{0}.$$

(Recall that the rows of A^\top are the columns of A .)

- Rearranging terms in the equation above gives the following system of p linear equations in p unknowns $\mathbf{x} = (x_1, \dots, x_p)$:

$$(A^\top A)\mathbf{x} = A^\top \mathbf{b}. \quad (4)$$

As we have argued in the first bullet above, this system is guaranteed to have a solution. But it is possible that it has more than one solution (and hence infinitely-many solutions).

The p linear equations in (4) are collectively called the normal equations. It turns out the normal equations have a unique solution precisely when $\text{rank}(A) = p$. This is implied by the following theorem.

Theorem 11. *For any matrix A , $\text{NS}(A) = \text{NS}(A^T A)$ and $\text{rank}(A) = \text{rank}(A^T A)$.*

Proof. We first show $\text{NS}(A) \subseteq \text{NS}(A^T A)$. If $A\mathbf{x} = \mathbf{0}$, then

$$(A^T A)\mathbf{x} = A^T(A\mathbf{x}) = A^T\mathbf{0} = \mathbf{0}.$$

Now we show $\text{NS}(A^T A) \subseteq \text{NS}(A)$. If $(A^T A)\mathbf{x} = \mathbf{0}$, then

$$\mathbf{x}^T(A^T A)\mathbf{x} = \mathbf{x}^T\mathbf{0} = 0.$$

But the left-hand side above can also be written as $(A\mathbf{x})^T(A\mathbf{x}) = \|A\mathbf{x}\|^2$, which is zero only if $A\mathbf{x} = \mathbf{0}$ by positive definiteness of the norm.

We conclude that $\text{NS}(A) = \text{NS}(A^T A)$. In particular, $\dim(\text{NS}(A)) = \dim(\text{NS}(A^T A))$. By the Dimension Theorem, $\text{rank}(A) = \text{rank}(A^T A)$. \square

If the $p \times p$ matrix $A^T A$ has rank p , then it is invertible (by the Invertibility Theorem), and in this case, the unique solution to (4) is given by the algebraic expression

$$\mathbf{x} = (A^T A)^{-1} A^T \mathbf{b},$$

and an expression for the orthogonal projection of \mathbf{b} is

$$\mathbf{b}_0 = A\mathbf{x} = A(A^T A)^{-1} A^T \mathbf{b}.$$

In this case, the orthoprojector for $\text{CS}(A)$ is given by

$$P = A(A^T A)^{-1} A^T.$$

But even if (4) has infinitely-many solutions, all of them yield the same (unique) vector $A\mathbf{x} = P\mathbf{b} \in \text{CS}(A)$. So every solution \mathbf{x} to (4) is a minimizer of the least squares approximation objective $\|A\mathbf{x} - \mathbf{b}\|^2$.

A Proofs of the Cauchy-Schwarz Inequality and Triangle Inequality

There are many proofs of the Cauchy-Schwarz Inequality. In the case of 2-vectors, it follows immediately from the fact that the cosine function has range $[-1, 1]$.

Proof of Theorem 2. Suppose either of \mathbf{u} or \mathbf{v} is the zero vector. Then the inequality is true since $\langle \mathbf{u}, \mathbf{v} \rangle = 0$. So we may assume that neither \mathbf{u} nor \mathbf{v} is the zero vector. Let a and b denote positive real numbers such that $ab = 1$. By the non-negativity of the norm and bilinearity of the inner product,

$$0 \leq \|a\mathbf{u} - b\mathbf{v}\|^2 = \langle a\mathbf{u} - b\mathbf{v}, a\mathbf{u} - b\mathbf{v} \rangle = a^2 \langle \mathbf{u}, \mathbf{u} \rangle - 2 \langle \mathbf{u}, \mathbf{v} \rangle + b^2 \langle \mathbf{v}, \mathbf{v} \rangle,$$

where the last step uses $ab = 1$. Rearranging terms and dividing by 2 gives

$$\langle \mathbf{u}, \mathbf{v} \rangle \leq \frac{a^2}{2} \langle \mathbf{u}, \mathbf{u} \rangle + \frac{b^2}{2} \langle \mathbf{v}, \mathbf{v} \rangle = \frac{a^2}{2} \|\mathbf{u}\|^2 + \frac{b^2}{2} \|\mathbf{v}\|^2.$$

Since this inequality is true for any positive numbers a and b with $ab = 1$, we can choose $a = \sqrt{\|\mathbf{v}\|/\|\mathbf{u}\|}$ and $b = \sqrt{\|\mathbf{u}\|/\|\mathbf{v}\|}$, so the right-hand side becomes $\|\mathbf{u}\|\|\mathbf{v}\|$. This proves the claimed inequality.

Now suppose $\langle \mathbf{u}, \mathbf{v} \rangle = \|\mathbf{u}\|\|\mathbf{v}\|$ and neither \mathbf{u} nor \mathbf{v} is the zero vector. Then the first displayed inequality above (with the prescribed choices of $a > 0$ and $b > 0$) must hold with equality:

$$0 = \|a\mathbf{u} - b\mathbf{v}\|^2.$$

Since only the zero vector has norm equal to 0, we conclude that $a\mathbf{u} = b\mathbf{v}$. So \mathbf{u} and \mathbf{v} are scalar multiples of each other. \square

The Triangle Inequality (Theorem 1) is a consequence of the Cauchy-Schwarz Inequality (Theorem 2).

Proof of Theorem 1. Bilinearity of the inner product and Theorem 2 imply

$$\begin{aligned} \|\mathbf{u} + \mathbf{v}\|^2 &= \langle \mathbf{u}, \mathbf{u} \rangle + 2 \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle \\ &\leq \|\mathbf{u}\|^2 + 2 \|\mathbf{u}\| \|\mathbf{v}\| + \|\mathbf{v}\|^2. \end{aligned}$$

The final right-hand side above is $(\|\mathbf{u}\| + \|\mathbf{v}\|)^2$. \square