CIS3990-002: Mathematics of Machine Learning Fall 2023

Lecture: Linear Algebra

Date: October 4th, 2023 *Author: Eric Wong*

1 Linear Algebra Basics

Most likely you are familiar with basic operations on matrices and vectors. For example, $A \in \mathbb{R}^{3\times 5}$ is a matrix of real numbers with 3 rows and 5 columns, while $b \in \mathbb{R}^3$ is a vector of 3 elements. These form the basis of a system that we call linear algebra, which has several main properties. Our goal in this module will be to learn about the the fundamental properties of linear systems, and generalize these properties to abstract vector spaces that are not necessarily in the field of real numbers.

- For $m, n \in \mathbb{N}$, a matrix A is a m, n tuple of elements a_{ij} where i denotes the row and j denotes the column.
- Addition: if $C = A + B$ and $A, B \in \mathbb{R}^{m\times n}$ then $c_{ij} = a_{ij} + b_{ij}$
- Product: If $C = AB$ and $A \in \mathbb{R}^{m\times k}$ and $B \in \mathbb{R}^{k\times n}$ then $c_{ij} = \sum_{l=1}^{k} a_{il} b_{lj}$ where $C \in \mathbb{R}^{m\times n}$
- Identity: $I_m \in \mathbb{R}^{m\times m}$ is an identity matrix when it is zero everywhere except the diagonal, i.e. $I_{ij} = \mathbf{1}[i = j]$
- Associativity: $(AB)C = A(BC)$
- Distributivity: $(A + B)C = AC + BC$, $A(C + D) = AC + AD$
- Multiplication with identity: $\forall A \in \mathbb{R}^{m\times n} : I_m A = A I_n = A$
- Inverse: Let $A \in \mathbb{R}^{n\times n}$. If $AB = I$ then $B = A^{-1}$ is the inverse of A
- Transpose: Let $A \in \mathbb{R}^{m\times n}$. The matrix $B = A^\top$ such that $b_{ij} = a_{ji}$ is called the transpose.
- Symmetric: $A \in \mathbb{R}^{n\times n}$ is symmetric if $A = A^\top$

We can also add scalars to the mix (single elements).

- Scalar multiplication: Let $\lambda \in \mathbb{R}$. Then, $\lambda A = K$ where $K_{ij} = \lambda a_{ij}$.
- Associativity: $(\lambda\phi)C = \lambda(\phi C)$. Actually, scalars can be moved around: $\lambda(BC) = (\lambda B)C = B(\lambda C) = (BC)\lambda$. Also, transpose doesn't affect matrices: $(\lambda C)^\top = C^\top \lambda = \lambda C^\top$
- Distributivity: $(\lambda + \phi)C = \lambda C + \phi C$ and $\lambda(B + C) = \lambda B + \lambda C$

One of the most common uses of matrices and vectors is to represent linear systems of equations in a compact form. I.e.

$$Ax = b$$

represents a series of linear equations, where each row of A is the coefficients for each variable x and the target scalar is the corresponding row in b.

2 Groups

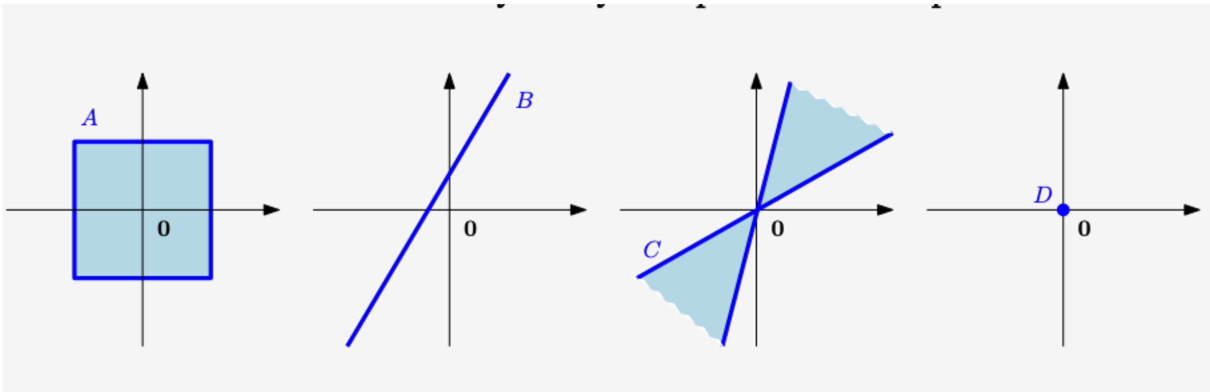
The space of matrices and vectors behaves *nicely*, in that it has these properties of associativity, distributivity, an identity and an inverse. Let's now generalize this structure.

- Groups: Let $\mathcal{G}$ be a set and an operation $\otimes : \mathcal{G} \times \mathcal{G} \to \mathcal{G}$ be defined on $\mathcal{G}$. Then $G = (\mathcal{G}, \otimes)$ is called a group if
  1. Closure: $\forall x, y \in \mathcal{G} : x \otimes y \in \mathcal{G}$
  2. Associativity: $\forall x, y, z \in \mathcal{G} : (x \otimes y) \otimes z = x \otimes (y \otimes z)$
  3. Neutral element: $\exists e \in \mathcal{G} \forall x \in \mathcal{G} : x \otimes e = e \otimes x = x$
  4. Inverse element: $\forall x \in \mathcal{G} \exists y \in \mathcal{G} : x \otimes y = y \otimes x = e$. We write x^{-1} to denote the inverse element of x. This does not always mean $\frac{1}{x}$ and is with respect to the operator $\otimes$.
  5. (Commutivity) If $f \forall x, y \in \mathcal{G} : x \otimes y = y \otimes x$ then $\mathcal{G}$ is an Abelian group
- Examples of Abelian groups: $(\mathcal{Z}, +), (\mathcal{R} \setminus \{0\})$
- Examples of not-groups: $(\mathcal{N} + 0, +), (\mathcal{Z}, \cdot), (\mathcal{R}, \cdot)$
- $(\mathcal{R}^n, +), (\mathcal{Z}^n, +)$ are Abelian if using component wise addition
- Matrices and addition: $(\mathcal{R}^{m \times n}, +)$ is Abelian with component-wise addition
- Matrices and multiplication: $(\mathcal{R}^{m \times n}, \cdot)$ is only a group if the inverse always exists
- General Linear Group: set of invertible matrices $A \in \mathcal{R}^{n \times n}$ is a group with respect to matrix multiplication, but is not Abelian (not commutative)

3 Vector Spaces

Groups have an operation with structure that stays within the group. This can be referred to as an *inner* operation (i.e. elementwise addition) as the operator stays within the group. We can also consider an *outer* operation which takes in an element outside of the group.

- Real-valued vector space $V = (\mathcal{V}, +, \cdot)$ is a set $\mathcal{V}$ with operations $+ : \mathcal{V} \times \mathcal{V} \to \mathcal{V}$ and $\cdot : \mathcal{R} \times \mathcal{V} \to \mathcal{V}$
- Distributivity:
  1. $\forall \lambda \in \mathbb{R}, \forall x, y \in \mathcal{V} : \lambda \odot (x + y) = \lambda \odot x + \lambda \odot y$
  2. $\forall \lambda, \psi \in \mathbb{R}, x \in \mathcal{V} : (\lambda + \psi) \odot x = \lambda \odot x + \psi \odot y$
- $x \in V$ are called vectors, the neutral element is 0, and the inner operator is vector addition while the outer operation is multiplcation by scalars.
- A subspace of a vector space is a vector space: if $\mathcal{U} \subset \mathcal{V}$ and $V = (\mathcal{V}, +, \odot)$ is a vector space, then if $U = (\mathcal{U}, +, \cdot)$ is a vector space we call it a subspace of V restricted to $\mathcal{U}$.



- Subspaces inherit properties from the higher space, including Abelian, distributivity, associativity, and neutral element. To show that U is a subspace, we need to show that $0 \in \mathcal{U}$ and U is closed with respect to both inner and outer operations (i.e. $\forall \lambda \forall x \in \mathcal{U} : \lambda x \in \mathcal{U}$ and $\forall x, y \in \mathcal{U} : x + y \in \mathcal{U}$).

- Example 2.12 from the textbook.

This structure gives us the nice properties we expect in linear algebra (i.e. we can do operations on vectors that result in more vectors).

4 Linear Independence, basis and rank

- Linear combination is a combination of scaled vectors:

$$v = \sum_i \lambda_i x_i$$

- If $x_i \in \mathbb{R}^d$ then we will typically abbreviate this as $\lambda^\top X$ where X is the matrix of elements stacked in each row

- If there exists λ such that $0 = \sum_i \lambda_i x_i$ with at least one $\lambda_i \neq 0$ then they are linearly dependent. If no such non-zero solution exists, they are linearly independent.

- Properties of linear independence
  - k vectors are either linearly dependent or independent
  - If a vector is 0 or if the same vector is repeated, they are dependent

- Let $V = (\mathcal{V}, +, \cdot)$ be a vector space and let $\mathcal{A} = \{x_1, \ldots, x_k\} \subset \mathcal{V}$. If every vector $v \in \mathcal{V}$ can be excpressed as a linear combination of $\mathcal{A}$, then this is a *generating set* of $\mathcal{V}$.

- The set of all linear combinations of vectors in $\mathcal{A}$ is the *span* of $\mathcal{A}$.

- A generating set $\mathcal{A}$ is *minimal* if there does not exist a smaller $\bar{\mathcal{A}} \subsetneq \mathcal{A}$ that spans V.

- Every independent generating sets of V is minimal and is called a *basis* of V.

- Let $\mathcal{B} \subseteq \mathcal{V}$, $\mathcal{B} \neq \emptyset$. The following statements are equivalent:
  - $\mathcal{B}$ is a basis
  - $\mathcal{B}$ is a minimal generating set
  - $\mathcal{B}$ is a maximally linearly independent set of vectors in V, i.e. adding any vector will make the set linearly dependent
  - Every vector $x \in V$ is a linear combination of vectors from $\mathcal{B}$, and every linear combination is unique:
$$x = \sum_i \lambda_i b_i = \sum_i \psi_i b_i \Rightarrow \lambda_i = \psi_i$$
- Example: Standard basis is $\mathcal{B} = \{e_1, \ldots, e_k\}$ where e_i is zero everywhere except for the ith position which is 1.
- Example:
$$\mathcal{B} = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right\}$$
- Every vector space has a basis $\mathcal{B}$. There can be multiple bases, i.e. they are not unique. However, they all have the same number of basis vectors.
- The *dimension* of V is the number of basis vectors of V, denoted as $\dim(V)$.
- A finite dimensional vector space is one where $\dim(V) < \infty$.
- Example of an infinite dimensional basis: suppose $\mathcal{V}$ is the set of all countably infinite vectors $v = (v_1, v_2, \ldots,)$. Then it has an infinite basis $\mathcal{B} = \{e_1, e_2, \ldots,\}$. Every vector can be written as $v = \sum_i \lambda_i e_i$ for some λ_i that exists for each v.
- When we go to function spaces, these will be infinite dimensional spaces.
- Example: The functions $e_n(\theta) = e^{2\pi i n\theta}$ is an (orthonormal) basis of the Hilbert space $L^2([0, 1])$ where $L^2([0, 1])$ is the space of functions on $[0, 1]$ for which the Lebesgue integral of the square of the absolute value is finite, i.e. $\int_X |f|^2 d\mu < \infty$
- The number of linearly independent columns of a matrix A equals the number of linearly independent rows and is called the rank of A, denoted as $\mathrm{rk}(A)$.
- Properties:
  - $\mathrm{rk}(A) = \mathrm{rk}(A^\top)$
  - The columns of A span a subspace $U \subseteq R^m$ with $\dim(U) = \mathrm{rk}(A)$. This subspace is called the image or range of A.
  - Similarly, the rows of A span a subspace $W \subseteq \mathbb{R}^n$ with $\dim(W) = \mathrm{rk}(A)$
  - For square matrices $A \in \mathbb{R}^{n \times n}$, A is invertible (regular) if and only if $\mathrm{rk}(A) = n$.
  - For all A, b the linear system $Ax = b$ can be solved if and only if $\mathrm{rk}(A) = \mathrm{rk}(A|b)$.
  - For $A \in \mathbb{R}^{m \times n}$, the subspace of solutions x such that $Ax = 0$ has rank $n - \mathrm{rk}(A)$. This subspace is called the kernel, or the null space.

- A matrix has full rank if $\text{rk}(A) = \min(m, n)$. Otherwise, it is rank deficient.

The goal of a basis is to provide a sense of structure to the vector space. We will now look at linear mappings, which are operations that preserve the structure of a vector space. This is analogous to the group operator, which preserves the structure of a group. Previously, we had operators $+$ and $\cdot$ for a vector space corresponding to elementwise addition and scalar multiplication. We will now look at operators between vector spaces that preserve this structure.

- Let V, W be two vector spaces. A linear mapping $\Phi : V \to W$ satisfies

$$\forall x, y \in V, \forall \lambda, \psi \in \mathbb{R} : \Phi(\lambda x + \psi y) = \lambda \Phi(x) + \psi \Phi(y)$$

  These are sometimes also called vector space homomorphism or linear transformation.

- In the vector space of $\mathbb{R}^n$, we can represent linear mappings as matrices.
- A mapping $\Phi : \mathcal{V} \to W$ on arbitrary sets $\mathcal{V}, \mathcal{W}$ is called:
  - Injective if $\forall x, y : \Phi(x) = \Phi(y) \Rightarrow x = y$ (different vectors map to different outputs)
  - Surjective if $\Phi(\mathcal{V}) = \mathcal{W}$ (all elements can be reached)
  - Bijective if it is both injective and surjective (operation can be undone)
- $\Phi : V \to W$ is an isomorphism if it is both linear and bijective
- $\Phi : V \to V$ is an endomorphism if it is linear
- $\Phi : V \to V$ is an automorphism if it is both linear and bijective
- $\text{id}_V : V \to V$ is the identity mapping, or automorphism.
- Theorem: Finite dimensional vector spaces V, W are isomorphic if and only if $\dim(V) = \dim(W)$ (Axler 2015).
- Intuitively, this means that vector spaces with the same dimension are the "same" in that you can transform from one to the other without losing anything. This means that we can treat the space of $\mathbb{R}^{m \times n}$ matrices as the same as $\mathbb{R}^{mn}$ vectors, as there is a linear bijective mapping from one to the other.
- More properties:
  - Let V, W, X be vector spaces. If $\Phi : V \to W$ and $\Psi : W \to X$ are linear mappings, then $\Psi \circ \Phi : V \to X$ is also linear.
  - If $\Phi : V \to W$ is an isomorphism, then $\Phi^{-1} : W \to V$ is an isomorphism.
  - If $\Phi : V \to W, \Psi : V \to W$ are linear, then $\Phi + \Psi$ and $\lambda \Phi$ are also linear.

The key point of the previous section is to say that any n dimensional vector space is isomorphic to $\mathbb{R}^n$. Therefore, any reasoning we can do in $\mathbb{R}^n$ applies to any finite dimensional vector space. So in the finite dimensional case, we only need to study $\mathbb{R}^n$, since everything that is finite can be reduce to $\mathbb{R}^n$. As an example, suppose

- Let $B = (b_1, \ldots, b_n)$ be an ordered basis of V. For any $x \in V$, let $x = \sum_i \alpha_i b_i$ be its unique linear combination. Then, we call $\alpha_1, \ldots, \alpha_n$ the coordinates of x.
- Think of a basis as definining a coordinate system.
- Typical coordinate system: standard basis $e_1, \ldots e_n$. The coefficients tell us how to linearly combine to obtain x. However, one could also use the basis $((1,0),(1,1))$ to span $\mathbb{R}^2$.
- Transformation matrix: Let V, W be vector spaces with bases B, C, and consider a linear mapping $\Phi : V \to W$. For $j \in \{1, \ldots, n\}$ let

  $$\Phi(b_j) = \sum_i \alpha_{ij} c_i$$

  be the unique representation of $\Phi(b_j)$ with respect to C. Then, if A is the matrix given by $A_{ij} = \alpha_{ij}$ then A is the transformation matrix of Φ with respect to B and C. This tells us how to go from one vector space to another but representented as a matrix.
- What this means is that *any linear mapping between finite dimension spaces can be represented with a matrix.* Just pick a basis for the domain and target, and compute the coefficients!
- If $\hat{x}$ is the coordinate vector of $x \in V$ and $\hat{y}$ is the coordinate vector of $y = \Phi(x) \in W$ then $\hat{y} = A_\Phi \hat{x}$ where A is the transformation matrix of Φ.

We now have a coordinate system for our vector spaces, which depends on a chosen basis $\mathcal{B}$. However, remember that the basis is not unique: there are multiple different possibly bases for a vector space. Depending on the basis, the resulting linear transformation could be easier or harder to work with. We will work towards characterizing what it means for a basis to be "nice". But in order to do so, we first we need to understand how to change between bases.

As an example, consider the linear transformation Φ with transformation matrix

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

in the standard basis. If instead of the standard basis, we use the basis $B = ([1,1],[1,-1])$ then the linear map Φ has transformation matrix

$$A = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$$

which is a digonal matrix (which is nice). We'll see how to get this in the next section.

- For a linear mapping $\Phi : V \to W$, consider two bases $B = (b_1, \ldots, b_n)$ and $\tilde{B} = (\tilde{b}_1, \ldots, \tilde{b}_n)$ on V and two bases $C = (c_1, \ldots, c_m)$ and $\tilde{C} = (\tilde{c}_1, \ldots, \tilde{c}_m)$ on W.
- Let $A_\Phi \in \mathbb{R}^{m \times n}$ be the transformation matrix of Φ with respect to B, C and let $\tilde{A}_\Phi$ be the transformation matrix of Φ with respect to $\tilde{B}, \tilde{C}$.
- How are A and $\tilde{A}$ related?

- Theorem: (Basis Change) The transformation matrix of $\tilde{A}_\Phi$ is given by

$$\tilde{A}_\Phi = T^{-1} A_\Phi S$$

where $S \in \mathbb{R}^{n \times n}$ is the transformation matrix of the id_V that maps $\tilde{B}$ onto B and $T \in \mathbb{R}^{m \times m}$ is the transformation matrix of id_W that maps coordinates with respect to $\tilde{C}$ to coordinates with respect to C.

- Proof: First, by definition of S we can write the $\tilde{b}_j$ as a sum of basis vectors b_i:

$$\tilde{b}_j = \sum_i s_{ij} b_i$$

Similarly, we can write $\tilde{c}_k$ as a combination of basis vectors of C:

$$\tilde{c}_k = \sum_l t_{lk} c_l$$

Then, S maps $\tilde{B}$ onto B and T maps $\tilde{C}$ onto C (the columns are the coordinate representation of $\tilde{b}_j$ and $\tilde{c}_k$ with respect to B and C). Now, re-express $\Phi(\tilde{b}_j)$ in two ways using these two bases. First using C:

$$\Phi(\tilde{b}_j) = \sum_{k=1}^{m} \tilde{a}_{kj} c_k = \sum_{k=1}^{m} \tilde{a}_{kj} \sum_{l=1}^{m} t_{lk} c_l = \sum_{l=1}^{m} c_l \sum_{k=1}^{m} \tilde{t}_{lk} a_{kj}$$

Then using B:

$$\Phi(\tilde{b}_j) = \Phi\left(\sum_{i=1}^{n} s_{ij} b_i\right) = \sum_{i=1}^{n} s_{ij} \Phi\left(b_i\right) = \sum_{i=1}^{n} s_{ij} \sum_{l=1}^{m} a_{li} c_l = \sum_{l=1}^{m} c_l \sum_{i=1}^{n} a_{li} s_{ij}$$

Therefore for all $j = 1, \ldots, n$ and al $l = 1, \ldots, m$ it follows that

$$\sum_{k=1}^{m} \tilde{t}_{lk} a_{kj} = \sum_{i=1}^{n} a_{li} s_{ij}$$

In matrix form, this is equivalent to

$$T\tilde{A} = AS$$

and therefore $\tilde{A} = T^{-1} AS$

- Aside: Why are S and T regular (invertible)? They are the matrix representation of the identity operator, which is an invertible operator.

- Two matrices $A, \tilde{A} \in \mathbb{R}^{m \times n}$ are *equivalent* if there exists regular matrices $S \in \mathbb{R}^{n \times n}, T \in \mathbb{R}^{m \times m}$ such that $\tilde{A} = T^{-1} AS$

- Two matrices $A, \tilde{A} \in \mathbb{R}^{n \times n}$ are *similar* if there exists a regular matrix $S \in \mathbb{R}^{n \times n}$ where $\tilde{A} = S^{-1} AS$

- Informally, this basis change can be seen as the following:

  - A maps $V \to W$ bases B to C

- $\tilde{A}$ maps $V \to W$ from bases $\tilde{B}$ to $\tilde{C}$
- S is the identity mapping from basis $\tilde{B}$ to B
- T is the identity mapping from basis $\tilde{C}$ to C
- Then, $\tilde{B} \to \tilde{C}$ can be rewritten as

$$\tilde{B} \to B \to C \to \tilde{C}$$

which reflects S, then A, then T^{-1}. Hence, $\tilde{A}x = T^{-1}(A(Sx))$

- To get the example from the start of this section: note that $S = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ (i.e. just horizontally stack the representation of the new basis in the old basis) and that

$$T^{-1} = S^{-1} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

Then,

$$\tilde{A} = T^{-1}AS = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$$

Now that we have a better understanding of bases and how to change between bases for linear mappings, we'll now cover our last fundamental concept for vector spaces: images and kernels.

- For $\Phi : V \to W$ the *kernel* or null space is $\ker(\Phi) = \phi^{-1}(0) = \{v \in V : \Phi(v) = 0\}$
- This is the set of vectors in V that map to 0
- The *image* or range is $\mathrm{Im}(\Phi) = \Phi(V)\{w \in W | \exists v \in V : \Phi(v) = w\}$
- This is the set of vectors in W that get mapped to
-
  - It is always true that $\Phi(0) = 0$ and therefore $0 \in \ker(\Phi)$, so the null space is never empty.
  - It is also always true that $\ker(\Phi) \subseteq V$ and $\mathrm{Im}(\Phi) \subseteq W$.
  - Φ is injective if and only if $\ker(\Phi) = \{0\}$
- Consider a linear mapping $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ with transformation matrix $A \in \mathbb{R}^{m \times n}$, so $\Phi(x) = Ax$
- let $A = [a_1, \ldots, a_n]$ be the columns of A. Then the image is the span of the columns (column space):

$$\mathrm{Im}(\Phi) = \{Ax : x \in \mathbb{R}^n = \sum_i x_i a_i : x_i \in \mathbb{R}\} = \mathrm{span}[a_1, \ldots, a_n] \subset mathbbA$$

- Then it follows that the rank of A is the dimension of the image, i.e. $\mathrm{rk}(A) = \dim(\mathrm{Im}(\Phi))$
- Rank Nullity Theorem: For vector spaces V, W and linear mapping $\Phi : V \to W$ it holds that

$$\dim(\ker(\Phi)) + \dim(\mathrm{Im}(\Phi)) = \dim(V)$$

also known as the fundamental theorem of linear mappings.

- Some immediately consequences:
  - If $\dim(\mathrm{Im}(\Phi)) < \dim(V)$ then $\ker(\Phi)$ is non-trivial
  - If A_Φ is the transformation matrix of Φ and $\dim(\mathrm{Im}(\Phi)) < \dim(V)$ then $Ax = 0$ has infinitely many solutions
  - If $\dim(V) = \dim(W)$ then Φ is injective if and only if it is surjective

The last part we will consider here is affine subspaces. These are subspaces that have a linear structure.

- Let V be a vector space $x_0 \in V$ and $U \subseteq V$ be a subspace. Then

$$L = x_0 + U = \{x_0 + u : u \in U\}$$

  is an affine subspace.

- Examples of affine subspaces: points, lines, planes...
- If $(b_1, \ldots, b_k)$ is an ordered basis of U then every element $x \in L$ is uniquely described as $x = x_0 + \sum_i \lambda_i^k b_i$
- In the same way that we can define linear mappings between vector spaces, we can also define affine mappings between affine subspaces.
- For two vector spaces V, W, linear mapping $\Phi : V \to W$ and $a \in W$, the mapping

$$\phi : V \to W$$

$$x \to a + \Phi(x)$$

  is an affine mapping from V to W, where a is the translation vector.

- Every affine mapping is the composition of a linear mapping Φ and a translation τ such that $\phi = \tau \odot \Phi$
- Composition $\phi' \odot \phi$ of affine operators is affine
- Affine operators preserve dimension and parellelism and other geometric structures

5 Inner Product Spaces

- A norm on a vector space V is the function

$$\|\cdot\| : V \to \mathbb{R}$$

  where we say $\|x\|$ is the norm of x. A norm must satisfy

  1. Absolutely homogeneous: $\|\lambda x\| = |\lambda|\|x\|$
  2. Triangle inequality: $\|x + y\| \ leq \|x\| + \|y\|$
  3. Positive definite: $\|x\| \geq 0$ and $\|x\| = 0 \Rightarrow x = 0$

- A norm represents the *size* of a vector
- Examples for $V = \mathbb{R}^d$:

  1. ℓ_1 norm $\|x\|_1 = \sum_i |x_i|$ (also called the Manhattan Norm),
  2. ℓ_2 norm $\|x\|_2 = \sqrt{\sum_i x_i^2} = \sqrt{x^T x}$ (also called the Euclidean distance, typically assumed as default),
  3. ℓ_0 norm $\|x\|_0 = \sum_i \nVdash[x_i \neq 0]$
  4. ℓ_∞ norm $\|x\|_\infty = \max_i |x_i|$

- Recall: a linear map is one that is linear with respect to addition and multiplication with a scalar, i.e. $\Phi(\lambda x + \psi y) = \lambda\Phi(x) + \psi\Phi(y)$
- A bi-linear map $\Omega : V \times V \to V$ is a function of two arguments that is linear in both arguments:

$$\Omega(\lambda x + \psi y, z) = \lambda\Omega(x, z) + \psi\Omega(y, z)$$

$$\Omega(x, \lambda y + \psi z) = \lambda\Omega(x, y) + \psi\Omega(x, z)$$

- A bi-linear map Ω is symmetric if

$$\Omega(x, y) = \Omega(y, x)$$

- A bi-linear map is positive definite if

$$\forall x \in V \setminus \{0\} : \Omega(x, x) > 0, \quad \Omega(0, 0) = 0$$

- A positive definite, symmetric bilinear map Ω is an inner product on V, where we often write $\Omega(x, y) = \langle x, y\rangle$
- $(V, \langle\cdot, \cdot\rangle$ is called an *inner product space.*
- Inner products capture the notion of *alignment* between two vectors x, y. This is analagous to the angle between vectors in Euclidean space.
- If $\langle x, y\rangle = x^T y$ then this is a Euclidean vector space
- Examples of inner products:

1. $\langle x, y \rangle = x^T y$
2. $\langle x, y \rangle = (x_1 y_2 + x_2 y_1) + 2x_2 y_2$

- Cauchy-Schwarz inequality:
$$|\langle x, y \rangle| \leq \|x\| \|y\|$$

With an inner product, we can define other useful constructs such as positive definite matrices.

- Let V be an inner product space with an ordered basis $B = (b_1, \ldots, b_n)$. Then, any vectors $x, y \in V$ can be written as a sum of basis vectors $x = \sum_i \psi_i b_i$ and $y = \sum_j \lambda_j b_j$ (so $\hat{x} = (\psi_1, \ldots, \psi_n)$ and $\hat{y} = (\lambda_1, \ldots, \lambda_n)$. Then,
$$\langle x, y \rangle = \left\langle \sum_i \psi_i b_i, \sum_j \lambda_j b_j \right\rangle = \sum_i \sum_j \psi_i \langle b_i, \rangle b_j \lambda_j = \hat{x}^T A \hat{y}$$
where $A_{ij} = \langle b_i, b_j \rangle$.
- Similar to how a linear map is uniquely determined by its transformation matrix in a given basis, an inner product is uniquely determined through A. Since an inner product is positive definite, it follows that $x^T A x > 0$ for all $x \in V \setminus \{0\}$.
- If $x^T A x > 0$ for all non-zero $x \in V$, then A is positive definite.
- If $x^T a x \geq 0$ for all non-zero $x \in V$, then A is positive semi-definite.
- Theorem: For a real-valued, finite dimensional vector space V with ordered basis B, $\langle \cdot, \cdot \rangle$ is an inner product if and only if there exists a symmetric positive definite matrix A such that $\langle x, y \rangle = \hat{x}^T A \hat{y}$
- Properties of symmetric positive definite matrices:
  1. $\ker(A) = \{0\}$. Since $x^\top A x > 0$ for all $x \neq 0$, it follows that $Ax \neq 0$ if $x \neq 0$.
  2. $\ker(A) = \{0\}$
  3. Diagonal entries must be positive since $e_i^\top A e_i > 0$
- Therefore, a PSD matrix defines an inner product with respect to a change of basis.

Inner products can do more than just capture the alignment of two vectors: they're a formalization of the notion of length and distance as well.

- A norm of a vector $\|x\|$ formalizes the size or the length of x.
- Any inner product induces a norm: $\|x\| = \sqrt{\langle x, x \rangle}$
- Can think of this induced norm as the alignment of a vector with itself
- Not all norms are defined by an inner product
- A norm then defines a distance between two vectors as the size of the difference:
$$d(x, y) = \|x - y\| = \sqrt{\langle x - y, x - y \rangle}$$

- Classic example: Euclidean distance
- Formally this mapping is called a *metric* if
  1. d is positive definite, i.e. $d(x, y) \geq 0$ for all $x, y \in V$ and $d(x, y) = 0 \Leftrightarrow x = y$
  2. d is symmetric, i.e. $d(x, y) = d(y, x)$ for all $x, y \in V$
  3. Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in V$
- Don't confuse $\langle x, y \rangle$ with $d(x, y)$ which have similar properties but behave oppositely

6 Orthogonality

Now that we have distances and similarity, we can now formalize a notion of dissimilarity. This is useful in defining what is "left" after finding some basis vectors while minimizing redundancies.

- Orthogonality: two vectors x, y are orthogonal if $\langle x, y \rangle = 0$. We write this as $x \perp y$. If $\|x\| = \|y\| = 1$ then we further say x, y are orthonormal.
- Orthogonality generalizes the idea of perpendicular to arbitrary norms that are not the dot product.
- Orthogonality depends on the inner product! $(1, 1)$ and $(-1, 1)$ are orthogonal with respect to the dot product, but not with respect to the inner product $\langle x, y \rangle = x^\top \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} y$
- Orthogonal matrix: a square matrix A is orthogonal if and only if its colums are orthonormal.
- This implies that $A^\top A = AA^\top = I$ and so $A^{-1} = A^\top$
- Orthogonal matrices do not change the length of a vector x:

$$\|Ax\|^2 = \|x\|^2$$

- Orthogonality is most often used to construct a basis. B is an orthonormal basis (ONB) if $\langle b_i, b_j \rangle = 0$ for $i \neq j$ and $\langle b_i, b_i \rangle = 1$
- Such a basis has minimal redundancy between basis vectors (with respect to the inner product) and is standardized to all have a norm of 1
- Orthogonal complement: Let V be a D dimensional vector space and $U \subseteq V$ be a M dimensional subspace. The orthogonal complement $U^\perp \subseteq V$ is a $D - M$ dimensional subspace such $U^\perp$ contains all vectors in V that are orthogonal to every vector in U:

$$U^\perp = \{v \in V | \forall u \in U : \langle u, v \rangle = 0\}$$

- Note that $U \cap U^\perp = \{0\}$. Therefore, every vector x can be decomposed into

$$x = \sum_m \lambda_{m=1}^M b_m + \sum_{j=1}^{D-M} \psi_j b_j^\perp$$

- Example: the tangent line of a plane in 3D space defines a subspace that is the orthogonal complement of the plane.
- Projection: Let $U \subseteq V$ be a subspace of V. A linear mapping $\pi : V \to U$ is a projection if $\pi^2 = \pi \odot \pi = \pi$.
- Recall that linear mappings can be uniquely represented with transformation matrices (after choosing a basis). Thus, projection matrices P satisfy $P^2 = P$.
- An orthogonal projection is a linear mapping that finds the "closest" point on the subspace:

$$\pi_U(x) = \arg\min_{u \in U} \|x - u\|$$

If $\pi_U(x)$ is the closest point to x, then the difference $\langle \pi_U(x) - x, u \rangle$ must be orthogonal to all $u \in U$. If U has a basis B, then $\langle \pi_U(x) - x, b_i \rangle = 0$.

- Aside: projections to general sets instead of subspaces can be defined as the closest point to a set U with respect to some norm.
- Re-expressing this as a set of coordinates λ so $\pi_U(x) = B\lambda$ gives us

$$\langle b_i, x - B\lambda \rangle = 0$$

for all i. This means that

$$B^\top (x - B\lambda) = 0 \Leftrightarrow B^\top x = B^\top B\lambda$$

Since B is a basis of U, the columns are linearly independent so $B^\top B$ is invertible. Thus,

$$\lambda = (B^\top B)^{-1} B^\top x$$

Plug this back in to get

$$\pi_U(x) = B\lambda = B(B^\top B)^{-1} B^\top x$$

- In one dimension, this simplifies to $\lambda = \frac{1}{\|b\|^2} b^\top x$
- If a basis is orthonormal, then this simplifies greatly to

$$\pi_U(x) = B\lambda = BB^\top x$$

where the coordinates are $\lambda = B^\top x$

- Projections let us reason about spaces without solving for an exact solution.
- Projection onto affine spaces: recall that an affine space is $L = x_0 + U$. We can subtract x_0, project onto U, and then add back in x_0 to get the orthogonal projection:

$$\pi_L(x) = x_0 + \pi_U(x - x_0)$$

Lastly, we'll show some examples of inner products on functions, which will reflect what we need later when we look at functional analysis.

- Inner products on non-finite spaces can be generalized to vectors with infinite entries (countably infinite) and also continuous valued (uncountably infinite). In the latter case this becomes an integral.

- Let $u, v : \mathbb{R} \to \mathbb{R}$ be two functions. One typical inner product is

$$\langle u, v \rangle = \int_a^b u(x)v(x)dx$$

- Example: $u = \sin(x)$ and $v = \cos(x)$ is an odd function, and therefore the integral from $-\pi$ to π is 0. Thus, sin and cos are orthogonal functions.

- The collection of functions $\{1, \cos(x), \cos(2x), \cos(3x), \ldots, \}$ is orthogonal when $(a, b) = (-\pi, \pi)$. This is a space of even and periodic functions, and projecting onto this space is the fundamental idea behind Fourier series.

7 Decompositions

We'll now explore how various decompositions in linear algebra reduce down to change of basis.

- With norms we have lengths, and with inner products we have similarity between vectors (i.e. angles). Lastly, we can introduce a notion of volume.
- Determinant $\det(A) : \mathbb{R}^{n\times n} \to \mathbb{R}$ is a measure of the volume of a hyper-parallelepiped formed from the columns of A
- $\det(T) = \prod_{i=1}^{T} T_{ii}$ where T is triangular
- Laplace expansion along column j:

$$\det(A) = \sum_{k=1}^{N} (-1)^{k+j} a_{kj} \det(A_{kj})$$

- A matrix $A \in \mathbb{R}^{n\times n}$ is invertible if and only if $\det(A) \neq 0$
- Intuition: if a matrix is not invertible, then some of its columns are linearly dependent and it spans a space less than N. Therefore, the volume in N dimensional space is 0. On the other hand, a volume of 0 implies that the columns span a subspace with dimension less than N and is therefore non-inverible.
- $\det(AB) = \det(A)\det(B)$
- $\det(A) = \det(A^\top)$
- $\det(A^{-1}) = \frac{1}{\det(A)}$ if A is invertible
- Similar matrices have the same determinant, i.e. if $\tilde{A} = S^{-1}AS$ then $\det(A) = det(A^{-1})$. Therefore, determinent does not depend on the basis.
- Multiplying a row by λ scales $\det(A)$ by λ, and $\det(\lambda A) = \lambda^N \det(A)$
- Swapping rows or columns changes the sign of $\det(A)$
- $\det(A) \neq 0$ if and only if $\mathrm{rk}(A) = N$
- The trace of a matrix $\mathrm{tr}(A) = \sum_i a_{ii}$ is the sum of the diagonal elements
- Properties of trace:
  1. $\mathrm{tr}(A + B) = \mathrm{tr}(A) + \mathrm{tr}(B)$
  2. $\mathrm{tr}(\alpha A) = \alpha \mathrm{tr}(A)$
  3. $\mathrm{tr}(I) = N$
  4. $\mathrm{tr}(AB) = \mathrm{tr}(BA)$
- Furthermore, the only operator that satisfies these properties is the trace

- Trace is invariant under cyclic permutations:

$$\mathrm{tr}(AKL) = \mathrm{tr}(KLA)$$

  Simpler:

$$\mathrm{tr}(xy^T) = \mathrm{tr}(y^T x) = y^T x$$

- The trace of a linear map is the trace of the linear operator:

$$\mathrm{tr}(\Phi) = \mathrm{tr}(A_\Phi)$$

- The trace is independent of any basis, since

$$\mathrm{tr}(B) = \mathrm{tr}(S^{-1}AS) = \mathrm{tr}(SS^{-1}A) = \mathrm{tr}(A)$$

With the determinant and the trace, we can now cover eigenvalues and eigenvectors.

- Let A be square. $\lambda \in \mathbb{R}$ is an eigenvalue and $x \neq 0$ is a corresponding eigenvector if $Ax = \lambda x$.
- Eigenvectors are directions where the transformation matrix A is a direct scaling of the direction by λ. In other words, transforming a vector with A simply scales the vector by λ.
- The following are equivalent:
  1. λ is an eigenvalue of A
  2. There exists x such that $Ax = \lambda x$, or equivalently $(A - \lambda I)x = 0$ can be solved for $x \neq 0$
  3. $\mathrm{rk}(A - \lambda I) < n$
  4. $\det(A - \lambda I) = 0$
- x, y are collinear if $x = cy$ for some c. All vectors collinear to an eigenvector are also eigenvectors, i.e.

$$A(cx) = cAx = c\lambda x = \lambda(cx)$$

- A and $A^\top$ have the same eigenvalues
- Similar matrices $\tilde{A}$ and A have the same eigenvalues. Therefore, a linear map Φ has the same eigenvalues regardless of the choice of basis.
- In total, eigenvalues, determinant, and trace are all invariant under basis change.
- Symmetric, positive definite matrices have positive, real eigenvalues
- Spectral Theorem: if A is symmetric, there exists an ONB of V of eigenvectors of A, and each eigenvalue is real:

$$A = PDP^\top$$

  where P contains the eigenvectors and D is diagonal of the eigenvalues
- Theorem. The determinant is $\det(A) = \prod_{i=1}^N \lambda_i$. You can think of this as the "area" of the parallelepiped after a change of basis to one that behaves like a "standard" basis.
- Theorem. The trace is $\mathrm{tr}(A) = \sum_{i=1}^N \lambda_i$

A diagonal matrix is very "nice". It has fast computation of determinants, powers, inverses, etc. An eigenvalue decomposition is a way to transform matrices into a diagonal form, and is a direct application of the change of basis. Let's explore this in more detail.

- Recall two matrices A, D are similar if $D = P^{-1}AP$ where P is invertible.
- Diagonalizable: we would like to have matrices A that are similar to diagonal matrices D. Then, this would mean that applying the linear map A is equivalent to scaling each coordinate in the new basis P.
- If $D = \text{diag}(\lambda)$, and let A be a square matrix. Then

  $$AP = PD$$

  if and only if λ are the eigenvalues and P are the eigenvectors. Why? By definition,

  $$AP = [Ap_1, \ldots, Ap_N]$$

  $$PD = \lambda_1 p_1, \ldots, \lambda_N p_N$$

- Therefore, $D = P^{-1}AP$ if and only if D are the eigenvalues and P are the eigenvalues (i.e. the diagonalization of the matrix is the eigenvalues and eigenvectors).
- Eigendecomposition: A square matrix A can be factored into

  $$A = PDP^{-1}$$

  where P is square and D is diagonal with eigenvalues of A if and only if the eigenvectors of A form a basis of V.
- Symmetric matrices can always be diagonalized. In particular, $AA^\top$ and $A^\top A$ are always diagonalizable.

Lastly, we'll summarize a few of the major decompositions and how these can be understood as finding a change of basis that results in a nice transformation matrix.

- The Eigendecomposition is a change of basis for square, symmetric matrices $A = U\Lambda U^T$, which results in a *diagonal* transformation matrix. The resulting basis is an *orthonormal* basis (the eigenvectors) with respect to the standard inner product $\langle x, y\rangle = x^T y$.
- The singular value decomposition is a change of basis for rectangular matrices $A = U\Sigma V^T$, which results in a *diagonal* transformation matrix. The resulting bases are *orthonormal* with respect to the standard inner product $\langle x, y\rangle = x^T y$
- The QR decomposition is a change of basis for rectangular matrices $A = QR$ which results in an *identity* transformation matrix. The resulting bases are unchanged for the target but orthonormal for the input with respect to the standard inner product.
- The Cholesky decomposition for symmetric, positive definite matrices is $A = LDL^T$, which can be viewed as a change of basis that results in a *diagonal* transformation matrix. In particular, the resulting basis is an orthogonal basis with respect to the matrix inner product $\langle x, y\rangle = x^T Ay$.

- Sinkhorn normal form for square positive matrices is $A = D_1 S D_2$ which can be viewed as a change of basis that results in a double stochastic transformation matrix S.
- See many more decompositions at https://en.wikipedia.org/wiki/Matrix_decomposition

As an example, PCA is often described as finding a basis that retains as much information as possible. This is formalized as finding basis directions that maximize the variance of the coordinates:

$$\max_b \mathbb{V}[b^\top x]$$

where $b^\top x$ is the 1D projection of x onto the coordinate b. PCA finds basis vectors b that maximize this variance by iteratively repeating this on the remaining orthogonal subspace.