‘Mathematics for Machine Learning’—Notes

Malcolm

Started 8 July 2024

Contents

Appendix A

Linear Algebra

A.1 Fundamentals

A.1.1 Groups

Given a set $\mathcal{G}$ and an operation $\otimes : \mathcal{G} \times \mathcal{G} \to \mathcal{G}$ defined on $\mathcal{G}$, then $G := (\mathcal{G}, \otimes)$ is called a *group* if the following hold:

1. *Closure* of $\mathcal{G}$ under $\otimes : \forall x, y \in \mathcal{G} : x \otimes y \in \mathcal{G}$
2. *Associativity*: $\forall x, y, z \in \mathcal{G} : (x \otimes y) \otimes z = x \otimes (y \otimes z)$
3. *Neutral element*: $\exists e \in \mathcal{G} \, \forall x \in \mathcal{G} : x \otimes e = x$ and $e \otimes x = x$
4. *Inverse element*: $\forall x \in \mathcal{G} \, \exists y \in \mathcal{G} : x \otimes y = e$ and $y \otimes x = e$, where e is the neutral element.

The neutral element in the depends on $\otimes$.

Abelian groups
If *Commutativity*: $\forall x, y \in \mathcal{G} : x \otimes y = y \otimes x$ holds, then $G = (\mathcal{G}, \otimes)$ is an *Abelian group*

Examples

- $(\mathbb{Z}, +)$ is an Abelian group
- $(\mathbb{N}_0, +)$ is not a group; although $(\mathbb{N}_0, +)$ possesses a neutral element (0), the inverse elements are missing.
- $(\mathbb{Z}, \cdot)$ is not a group; although $(\mathbb{Z}, \cdot)$ contains the neutral element (1), the inverse elements for any $z \in \mathbb{Z}$, $z \neq \pm 1$, are missing

- $(\mathbb{R}^n, +), (\mathbb{Z}^n, +), n \in \mathbb{N}$ are Abelian if $+$ is defined componentwise, i.e.,

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n)$$

  Where $(-x_1, \ldots, -x_n)$ is the inverse element and $e = (0, \ldots, 0)$ is the neutral element

A.1.2 Vector Spaces

A real-valued vector space $V = (\mathcal{V}, +, \cdot)$ is a set $\mathcal{V}$ with two operations

$$+ : \mathcal{V} \times \mathcal{V} \to \mathcal{V}$$
$$\cdot : \mathbb{R} \times \mathcal{V} \to \mathcal{V}$$

(the first operation $+$ is an inner operation (mappings that only operate on elements in $\mathcal{G}$); the other operation is the multiplication of a vector $x \in \mathcal{G}$ by a scalar $\lambda \in \mathbb{R}$. We can think of the inner operation as a form of addition, and the outer operation as a form of scaling.)

where

1. $(\mathcal{V}, +)$ is an Abelian group
2. Distributivity:
   - $\forall \lambda \in \mathbb{R}, x, y \in \mathcal{V} : \lambda \cdot (x + y) = \lambda \cdot x + \lambda \cdot y$
   - $\forall \lambda, \psi \in \mathbb{R}, x \in \mathcal{V} : (\lambda + \psi) \cdot x = \lambda \cdot x + \psi \cdot x$
3. Associativity (outer operation): $\forall \lambda, \psi \in \mathbb{R}, x \in \mathcal{V} : \lambda \cdot (\psi \cdot x) = (\lambda \cdot \psi) \cdot x$
4. Neutral element with respect to the outer operation: $\forall x \in \mathcal{V} : 1 \cdot x = x$

(the neutral element of $(\mathcal{V}.+)$ is the zero vector $\mathbf{0} = [0, \ldots, 0]^T$) Note that the "vector multiplication" $ab, a, b \in \mathbb{R}^n$, is not defined only the following multiplications for vectors are defined: $ab^T \in \mathbb{R}^{n \times n}$ (outer product), $a^T b \in \mathbb{R}$ (inner/scalar/dot product).

Example:
$\mathcal{V} = \mathbb{R}^{m \times n}, m, n \in \mathbb{N}$ is a vector space with

- Addition: $A + B = \begin{bmatrix} a_{11} + b_{11} & \cdots & a_{1n} + b_{1n} \\ \vdots & & \vdots \\ a_{m1} + b_{m1} & \cdots & a_{mn} + b_{mn} \end{bmatrix}$ defined elementwise for all $A, B \in \mathcal{V}$.
- Multiplication by scalars: $\lambda A = \begin{bmatrix} \lambda a_{11} & \cdots & \lambda a_{1n} \\ \vdots & & \vdots \\ \lambda a_{m1} & \cdots & \lambda a_{mn} \end{bmatrix}$

A.1.3 Vector Subspaces

Let $V = (\mathcal{V}, +, \cdot)$ be a vector space and $\mathcal{U} \subseteq \mathcal{V}, \mathcal{U} \neq \emptyset$. Then $U = (\mathcal{U}, +, \cdot)$ is called a *vector subspace/linear subspace* of V if U is a vector space with the vector space operations $+$ and $\cdot$ restricted to $\mathcal{U} \times \mathcal{U}$ and $\mathbb{R} \times \mathcal{U}$. We write $U \subseteq V$ to denote a subspace U of V.

Properties
If $\mathcal{U} \subseteq \mathcal{V}$ is a vector space, then U naturally inherits many properties from V, since they hold for all $x \in V$, and thus $x \in \mathcal{U} \in \mathcal{V}$.

These include the *Abelian group properties, distributivity, associativity, and the neutral element*. To determine whether $(\mathcal{U}, +, \cdot)$ is a valid subspace of V need to show

1. $\mathcal{U} \neq \emptyset$, in particular: $\mathbf{0} \in \mathcal{U}$
2. Closure of U:
   (a) With respect to the outer operation: $\forall \lambda \in \mathbb{R} \, \forall x \in \mathcal{U} : \lambda x \in \mathcal{U}$
   (b) With respect to the inner operation: $\forall x, y \in \mathcal{U} : x + y \in \mathcal{U}$

A.1.4 Are linear combinations of linearly independent vectors also linearly independent?

Consider a vector space V with k linearly independent vectors $b_1 \dots b_k$; consider m linear combinations of these vectors:

$$\begin{aligned} x_1 &= \sum_{i=1}^{k} \lambda_{i1} b_i \\ &\vdots \\ x_m &= \sum_{i=1}^{k} \lambda_{im} b_i \end{aligned}$$

Defining $B = [b_1 \dots b_k]$ as a matrix with columns being the linearly independent vectors $b_1 \dots b_k$, we can write

$$x_j = B\lambda_j, \quad \lambda_j = \begin{bmatrix} \lambda_{1j} \\ \vdots \\ \lambda_{kj} \end{bmatrix}, \quad j = 1, \dots, m$$

We want to test whether $x_1, \dots, x_m$ are linearly independent. Using the general approach (whether $\sum_{j=1}^{m} \psi_j x_j = 0$ has nontrivial solutions):

$$\sum_{j=1}^{m} \psi_j x_j = \sum_{j=1}^{m} \psi_j B\lambda_j = B \sum_{j=1}^{m} \psi_j \lambda_j$$

This means that $\{x_1, \dots, x_m\}$ are linearly independent if and only if the column vectors $\{\lambda_1, \dots, \lambda_m\}$ are linearly independent.

A.1.5 Generating Set, Basis, Span

Generating set, Span
Consider a vector space $V = (\mathcal{V}, +, \cdot)$ and a set of vectors $\mathcal{A} = \{x_1, \ldots, x_k\} \subseteq \mathcal{V}$. If every vector $v \in \mathcal{V}$ can be expressed as a linear combination of $x_1, \ldots, x_k$, $\mathcal{A}$ is called a *generating set* of V.

The set of all linear combinations of vectors in $\mathcal{A}$ is called the *span* of $\mathcal{A}$. Where $\mathcal{A}$ spans the vector space V, we write $V = \text{span}[\mathcal{A}]$ or $V = \text{span}[x_1, \ldots, x_k]$.

Intuitively, every vector in a vector (sub)space can be represented as a linear combination of the vectors in the generating set that spans that vector (sub)space.

Basis
Consider a vector space $V = (\mathcal{V}, +, \cdot)$ and $\mathcal{A} \subseteq \mathcal{V}$. A generating set $\mathcal{A}$ of V is called *minimal* if there exists no smaller set $\tilde{\mathcal{A}} \subseteq \mathcal{A} \subseteq \mathcal{V}$ that spans V. Every linearly independent generating set of V is minimal and is called the *basis* of V.

For intuition, let $V = (\mathcal{V}, +, \cdot)$ be a vector space and $\mathcal{B} \subseteq \mathcal{V}, \mathcal{B} \neq \emptyset$. Then the following statements are equivalent:

- $\mathcal{B}$ is a basis of V
- $\mathcal{B}$ is a minimal generating set
- $\mathcal{B}$ is a maximal linearly independent set of vectors in V (adding any other vector in V to this set would make it linearly dependent).
- Every vector $x \in V$ is a linear combination of vectors from B, and every linear combination is unique; with

  $$x = \sum_{i=1}^{k} \lambda b_i = \sum_{i=1}^{k} \psi b_i$$

  and $\lambda_i, \psi_i \in \mathbb{R}$ it follows that $\lambda_i = \psi_i$, $i = 1 \ldots k$.

A.1.6 Dimensionality (finding a basis)

Dimensionality
Considering finite-dimensional vector spaces V, the *dimension* of V is the number of basis vectors of V, written as $\dim V$. If $U \subseteq V$ is a subspace of V, then $\dim(U) \leq \dim(V)$ and $\dim(U) = \dim(V)$ if and only if $U = V$. Finding the dimension of a vector space is therefore the same as finding a basis.

Determining a Basis
The basis of a subspace $U = \text{span}[x_1, \ldots, x_m] \subseteq \mathbb{R}^n$ can be found using the protocol:

1. Write the spanning vectors as columns of a matrix A
2. Determine the row-echelon form of A
3. The spanning vectors associated with the pivot columns are a basis of U

Consider an example: for a vector subspace $U \subseteq \mathbb{R}^5$, spanned by the vectors

$$x_1 = \begin{bmatrix} 1 \\ 2 \\ -1 \\ -1 \\ -1 \end{bmatrix}, \quad x_2 = \begin{bmatrix} 2 \\ -1 \\ 1 \\ 2 \\ -2 \end{bmatrix}, \quad x_3 = \begin{bmatrix} 3 \\ -4 \\ 3 \\ 5 \\ -3 \end{bmatrix}, \quad x_4 = \begin{bmatrix} -1 \\ 8 \\ -5 \\ -6 \\ 1 \end{bmatrix}, \in \mathbb{R}^5$$

in order to determine which vectors $x_1 \ldots x_4$ are a basis for U, we need to check whether they are linearly independent; we need to solve

$$\sum_{i=1}^{4} \lambda_i x_i = \mathbf{0}$$

Notice this can be viewed as solving the system represented by the augmented matrix

$$\left[\begin{array}{cccc|c} \vdots & \vdots & \vdots & \vdots & \vdots \\ x_1 & x_2 & x_3 & x_4 & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{array}\right]$$

(thus the 'augmented' part of the matrix can simply be ignored since gaussian elimination won't alter it)
(next page)

Using gaussian elimination to obtain the row-echelon form:

$$\begin{bmatrix} 1 & 2 & 3 & -1 \\ 2 & -1 & -4 & 8 \\ -1 & 1 & 3 & -5 \\ -1 & 2 & 5 & -6 \\ -1 & -2 & -3 & 1 \end{bmatrix} \rightsquigarrow \cdots \rightsquigarrow \begin{bmatrix} 1 & 2 & 3 & -1 \\ 0 & 1 & 2 & -2 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The non-pivot column corresponds to the vector in the generating set that can be expressed as a linear combination of the other vectors in that generating set. The pivot columns indicate which set of vectors are linearly independent.

From the pivot columns of the row-echelon form we see that x_1, x_2, x_4 are linearly independent. Therefore, $\{x_1, x_2, x_4\}$ is a basis of U.

A.1.7 Rank

The number of linearly independent columns of matrix $A \in \mathbb{R}^{m \times n}$ equals the number of linearly independent rows and is called the *rank* of A, denoted by $\text{rk}(A)$.

Properties

- $\text{rk}(A)$=$\text{rk}(A^T)$; the column rank equals the row rank.
- The columns of $A \in \mathbb{R}^{m \times n}$ span a subspace $U \subseteq \mathbb{R}^m$ (the number of rows determine the dimension the vector exists in) with $\dim(U) = \text{rk}(A)$ this subspace is the *image/range*, with basis found through Gaussian elimination of A to identify the pivot columns.
- The rows of $A \in \mathbb{R}^{m \times n}$ span a subspace $W \subseteq \mathbb{R}^n$ with $\dim(W) = \text{rk}(A)$. A basis for W can be found by Gaussian elimination of A^T.
- For all $A \in \mathbb{R}^{n \times n}$ it holds that A is regular/invertible if and only if $\text{rk}(A) = n$ (full rank).
- For all $A \in \mathbb{R}^{m \times n}$ and all $b \in \mathbb{R}^m$ it holds that the linear equation system $Ax = b$ can be solved if and only if $\text{rk}(A)$=$\text{rk}(A|b)$, where $A|b$ denotes the augmented system (otherwise there will be a contradiction).
- For $A \in \mathbb{R}^{m \times n}$ the subspace of solutions for $Ax = \mathbf{0}$ possesses dimension $n - \text{rk}(A)$. This subspace is the *kernel/nullspace*.
- A matrix $A \in \mathbb{R}^{m \times n}$ has *full rank* if its rank is equal to the largest possible rank for a matrix of the same dimensions; meaning $\text{rk}(A) = \min(m, n)$. A matrix is *rank deficient* if it is not full rank.

A.1.8 Linear Mappings

Consider two real vector spaces V, W. A mapping $\Phi : V \to W$ preserves the structure of the vector space if

$$\Phi(x + y) = \Phi(x) + \Phi(y)$$
$$\Phi(\lambda x) = \lambda\Phi(x)$$

for all $x, y \in V$ and $\lambda \in \mathbb{R}$.

For vector spaces V, W, a mapping $\Phi : V \to W$ is called a *linear mapping/vector space homomorphism/linear transformation* if

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

(*Homo-* meaning 'same' and *-morph* meaning 'form' or 'shape) Linear mappings can be represented by matrices.

Consider a mapping $\Phi : \mathcal{V} \to \mathcal{W}$, where $\mathcal{V}, \mathcal{W}$ can be arbitrary sets. Then Φ is called

- *Injective* if $\forall x, y \in \mathcal{V} : \Phi(x) = \Phi(y) \implies x = y$.
- *Surjective* if $\Phi(\mathcal{V}) = \mathcal{W}$.
- *Bijective* if it is injective and surjective.

Intuitively, if Φ is surjective, then every element in $\mathcal{W}$ can be 'reached' from $\mathcal{V}$. A bijective Φ can be 'undone', meaning there exists a mapping $\Psi : \mathcal{W} \to \mathcal{V}$ so that $\Psi \circ \Phi(x) = x$. This mapping Ψ can then be called the inverse of Φ and be denoted by Φ^{-1}.

Now we define special cases of linear mappings between vector spaces V and W:

- *Isomorphism*: $\Phi : V \to W$ linear and bijective
- *Endomorphism*: $\Phi : V \to V$ linear
- *Automorphism*: $\Phi : V \to V$ linear and bijective
- We define $\mathrm{id}_V : V \to V, x \mapsto x$ as the *identiy mapping/identity automorphism.*

(next page)

Example: Homomorphism
The mapping $\Phi : \mathbb{R}^2 \to \mathbb{C}, \Phi(x) = x_1 + ix_2$, is a homomorphism since

$$\Phi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}\right) = (x_1 + y_1) + i(x_2 + y_2) = x_1 + ix_2 + y_1 + iy_2$$
$$= \Phi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) + \Phi\left(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}\right)$$

and

$$\Phi\left(\lambda \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \lambda x_1 + \lambda i x_2 = \lambda(x_1 + ix_2) = \lambda\Phi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right)$$

Notice that this justifies why complex numbers can be represented as tuples in $\mathbb{R}^2$.

Theorem: *Finite-dimensional vector spaces V and W are isomorphic if and only if* $\dim(V) = \dim(W)$.

This theorem states that there exists a linear, bijective mapping between two vector spaces of the same dimension. Intuitively this means that vector spaces of the same dimension are kind of the same thing, since they can be transformed into each other without incurring any loss.

A.1.9 Matrix Representation of Linear Mappings

Ordered Basis

Any n-dimensional vector space is isomorphic to $\mathbb{R}^n$ (A.1.8). Consider a basis $\boldsymbol{\{b_1 \dots b_n\}}$ of an n-dimensional vector space V; in the following the order of the basis vectors will be important, so we write

$$B = \boldsymbol{(b_1 \dots b_n)}$$

and call this n-tuple an *ordered basis* of V.
(Just a note on notation: $B = \boldsymbol{(b_1 \dots b_n)}$ is an ordered basis, $B = \boldsymbol{\{b_1 \dots b_n\}}$ is an (unordered) basis, and $B = \boldsymbol{[b_1 \dots b_n]}$ is a matrix whose columns are the vectors $\boldsymbol{b_1 \dots b_n}$.)

Coordinates

Consider a vector space V and an ordered basis $B = \boldsymbol{(b_1 \dots b_n)}$ of V. For any $x \in V$ we have a unique representation (through a linear combination)

$$x = \alpha_1 \boldsymbol{b_1} + \ldots + \alpha_n \boldsymbol{b_n}$$

of x with respect to B. Then $\alpha_1 \dots \alpha_n$ are the *coordinates* of x with respect to B, and the vector

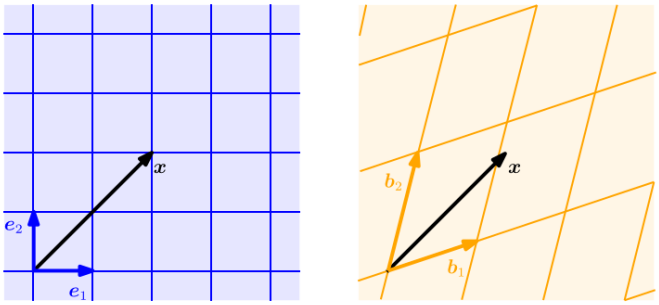
$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} \in \mathbb{R}^n$$

is the *coordinate vector/coordinate representation* of x with respect to the ordered basis B.

Intuitively, a basis effectively defines a coordinate system. One is usually familiar with the canonical basis vectors $\boldsymbol{e_1, e_2}$; notice that a vector in this coordinate system $x \in \mathbb{R}^2$ can also be seen as a representation describing how to linearly combine $\boldsymbol{e_1, e_2}$ to obtain x.

However, any basis of $\mathbb{R}^2$ defines a valid coordinate system; the same vector x from before may therefore have a different representation in the $\boldsymbol{(b_1, b_2)}$ basis. (next page)

Example
Consider how the same vector can be represented differently in two different basis representations.



In the above figure, the coordinates of x with respect to the standard basis is $[2,2]^T$, but with respect to the basis $(\boldsymbol{b_1}, \boldsymbol{b_2})$ the same vector x is represented as $[1.09, 0.72]^T$, where $x = 1.09\boldsymbol{b_1} + 0.72\boldsymbol{b_2}$ instead of $x = 2\boldsymbol{e_1} + 2\boldsymbol{e_2}$.

Another Example
The geometric vector $x \in \mathbb{R}^2$ with coordinates $[2,3]^T$ with respect to the standard basis $\boldsymbol{e_1}, \boldsymbol{e_2}$ of $\mathbb{R}^2$ can be written as $x = 2\boldsymbol{e_1} + 3\boldsymbol{e_2}$. However we don't necessarily have to use the standard basis to represent it; if we use the basis $\boldsymbol{b_1} = [1,-1]^T, \boldsymbol{b_2} = [1,1]^T$ we obtain the coordinates $\frac{1}{2}[-1,5]^T$ to represent the same vector with respect to $(\boldsymbol{b_1}, \boldsymbol{b_2})$; $x = -1/2 \cdot \boldsymbol{b_1} + 5/2 \cdot \boldsymbol{b_2}$.

Remark. For an n-dimensional space V and an ordered basis B of V, the mapping $\Phi : \mathbb{R}^n \to V, \Phi(e_i) = b_i, i = 1 \dots n$, is linear (and because of (A.1.8) an isomorphism), where $(\boldsymbol{e_1} \dots \boldsymbol{e_n})$ is the standard basis of $\mathbb{R}^n$.

Note that a *mapping* and a *change of basis* (like the examples above) are different.
(next page)

Transformation Matrix
Consider vector spaces V, W with corresponding (ordered) bases $B = (\boldsymbol{b_1} \dots \boldsymbol{b_n})$ and $C = (\boldsymbol{c_1} \dots \boldsymbol{c_m})$. Now consider a linear mapping $\Phi : V \to W$. For $j \in \{1, \dots, n\}$,

$$\Phi(\boldsymbol{b_j}) = \alpha_{1j}\boldsymbol{c_1} + \dots + \alpha_{mj}\boldsymbol{c_m} = \sum_{i=1}^{m} \alpha_{ij}\boldsymbol{c_i}$$

is the unique representation of $\Phi(\boldsymbol{b_j})$ with respect to the ordered basis C. We call this $m \times n$-matrix A_Φ, whose elements are given by

$$A_\Phi(i, j) = \alpha_{ij}$$

the *transformation matrix* of Φ (with respect to the ordered bases B of V and C of W).

The coordinates of $\Phi(\boldsymbol{b_j})$ with respect to the ordered basis C of W are the j-th column of A_Φ. If $\hat{\boldsymbol{x}}$ is the coordinate vector of $x \in V$ with respect to B and $\hat{\boldsymbol{y}}$ the coordinate vector of $y = \Phi(x) \in W$ with respect to C, then

$$\hat{\boldsymbol{y}} = \boldsymbol{A_\Phi}\hat{\boldsymbol{x}}$$

the transformation matrix can be used to map coordinates with respect to an ordered basis in V to coordinates with respect to an ordered basis in W.

Intuition
Consider coordinate vector $[a, b, c, \dots]^T$ in V with basis B (meaning the vector in V is expressed as $a\boldsymbol{b_1} + b\boldsymbol{b_2} + \dots$). Now consider applying a linear mapping $\Phi : V \to W$, where W is has basis C, to V:

$$\Phi(a\boldsymbol{b_1} + b\boldsymbol{b_2} + \dots) = a\Phi(\boldsymbol{b_1}) + b\Phi(\boldsymbol{b_2}) + \dots$$

we express each basis vector of B in terms of C; for $j \in \{1, \dots, n\}$,

$$\Phi(\boldsymbol{b_j}) = \alpha_{1j}\boldsymbol{c_1} + \dots + \alpha_{mj}\boldsymbol{c_m}$$

so

$$\begin{aligned} & a\Phi(\boldsymbol{b_1}) + b\Phi(\boldsymbol{b_2}) + \dots \\ =& a(\alpha_{11}\boldsymbol{c_1} + \dots + \alpha_{m1}\boldsymbol{c_m}) + b(\alpha_{12}\boldsymbol{c_1} + \dots + \alpha_{m2}\boldsymbol{c_m}) + \dots \\ =& \underbrace{(a\alpha_{11} + b\alpha 12 + \dots)}_{\text{first coordinate}}\boldsymbol{c_1} + (a\alpha_{m1} + b\alpha m2)\boldsymbol{c_m} + \dots \end{aligned}$$

which leads to our compact representation $\hat{\boldsymbol{y}} = \boldsymbol{A_\Phi}\hat{\boldsymbol{x}}$.

A.1.10 Basis Change

Intuition
Consider two ordered bases of V:

$$B = (\boldsymbol{b_1}, \ldots, \boldsymbol{b_n}), \quad \tilde{B} = (\tilde{\boldsymbol{b_1}}, \ldots, \tilde{\boldsymbol{b_n}})$$

and two ordered bases of W:

$$C = (\boldsymbol{c_1}, \ldots, \boldsymbol{c_m}), \quad \tilde{C} = (\tilde{\boldsymbol{c_1}}, \ldots, \tilde{\boldsymbol{c_m}})$$

We define $\boldsymbol{A_\Phi} \in \mathbb{R}^{m \times n}$ be the transformation matrix of the linear mapping $\Phi : V \to W$ with respect to bases B and C, and $\tilde{\boldsymbol{A}}_{\boldsymbol{\Phi}} \in \mathbb{R}^{m \times n} \in \mathbb{R}^{m \times n}$ as the corresponding transformation mapping with respect to $\tilde{B}$ and $\tilde{C}$. The idea is that by changing the basis and correspondingly the representation of vectors, the transformation matrix with respect to this new basis can have a particularly simple form allowing for straightforward computation.

Example
Consider a transformation matrix

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

with respect to the canonical basis in $\mathbb{R}^2$. If we define a new basis

$$B = \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right)$$

we obtain a diagonal transformation matrix

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

with respect to B, which is easier to work with than $\boldsymbol{A}$.

We want mappings that transform coordinate vectors with respect to one basis into coordinate vectors with respect to a different basis.
(next page)

Theorem (Basis Change)
For a linear mapping $\Phi : V \to W$, ordered bases

$$B = (\boldsymbol{b_1}, \ldots, \boldsymbol{b_n}), \quad \tilde{B} = (\tilde{\boldsymbol{b_1}}, \ldots, \tilde{\boldsymbol{b_n}})$$

of V and

$$C = (\boldsymbol{c_1}, \ldots, \boldsymbol{c_m}), \quad \tilde{C} = (\tilde{\boldsymbol{c_1}}, \ldots, \tilde{\boldsymbol{c_m}})$$

of W, and a transformation matrix $\boldsymbol{A_\Phi}$ of Φ with respect to B and C, the corresponding transformation matrix $\tilde{\boldsymbol{A}}_{\boldsymbol{\Phi}}$ with respect to the bases $\tilde{B}$ and $\tilde{C}$ is given as

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

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

Proof
We can write the vectors of the new basis $\tilde{B}$ of V as a linear combination of the basis vectors of B, such that

$$\tilde{\boldsymbol{b_j}} = s_{1j}\boldsymbol{b_1} + \ldots + s_{nj}\boldsymbol{b_n} = \sum_{i=1}^{n} s_{ij}\boldsymbol{b_i}, \quad j = 1, \ldots, n$$

Similarly, we write the basis vectors of $\tilde{C}$ of W as a linear combination of the basis vectors of C, which yields

$$\tilde{\boldsymbol{c_k}} = t_{1k}\boldsymbol{c_1} + \ldots + t_{mk}\boldsymbol{c_m} = \sum_{l=1}^{m} t_{lk}\boldsymbol{c_l}, \quad k = 1, \ldots, m$$

With that we we can define $\boldsymbol{S} = ((s_{ij})) \in \mathbb{R}^{n \times n}$ as the transformation matrix that maps coordinates with respect to $\tilde{B}$ onto coordinates with respect to B and $\boldsymbol{T} = ((t_{lk})) \in \mathbb{R}^{m \times m}$ as the transformation matrix that maps coordinates with respect to $\tilde{C}$ onto coordinates with respect to C.

(The jth column of $\boldsymbol{S}$ is the coordinate representation of $\tilde{\boldsymbol{b_j}}$ with respect to B; the kth column of $\boldsymbol{T}$ is the coordinate representation of $\tilde{\boldsymbol{c_k}}$ with respect to C. Note that both $\boldsymbol{S}$ and $\boldsymbol{T}$ are regular (A.1.4))
(next page)

The mapping $\Phi(\tilde{\boldsymbol{b_j}})$ can be written as

$$\Phi(\tilde{\boldsymbol{b_j}}) = \sum_{k=1}^{m} \tilde{a_{kj}} \tilde{\boldsymbol{c_k}} = \sum_{k=1}^{m} \tilde{a_{kj}} \sum_{l=1}^{m} t_{lk} \boldsymbol{c_l} = \sum_{l=1}^{m} \left(\sum_{k=1}^{m} t_{lk} \tilde{a_{kj}} \right) \boldsymbol{c_l}$$

here we've written each mapped basis vector of $\tilde{B}$ in terms of bases of $\tilde{C}$ (which gives us transformation matrix $\tilde{\boldsymbol{A}}_{\boldsymbol{\Phi}}$). Then we've written each basis vector of $\tilde{C}$ in terms of bases of C (this was shown above too, it gives us the transformation matrix $\boldsymbol{T}$).

Now notice we can also write the mapping as

$$\begin{aligned} \Phi(\tilde{\boldsymbol{b_j}}) &= \Phi \left(\sum_{i=1}^{n} s_{ij} \boldsymbol{b_i} \right) = \sum_{i=1}^{n} s_{ij} \Phi(\boldsymbol{b_i}) = \sum_{i=1}^{n} s_{ij} \sum_{l=1}^{m} a_{li} c_l \\ &= \sum_{l=1}^{m} \left(\sum_{i=1}^{n} a_{li} s_{ij} \right) c_l \end{aligned}$$

We write each basis of $\tilde{B}$ in terms of bases of B (this gives us $\boldsymbol{S}$), and then each basis of B as bases of C (giving us $\boldsymbol{A}_{\boldsymbol{\Phi}}$); the second step comes from the linearity of Φ (A.1.8).

It therefore follows that for all $j = 1, \ldots, n$ and $l = 1, \ldots, m$ that

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

and therefore

$$\boldsymbol{T} \tilde{\boldsymbol{A}}_{\boldsymbol{\Phi}} = \boldsymbol{A}_{\boldsymbol{\Phi}} \boldsymbol{S} \in \mathbb{R}^{m \times n}$$

such that

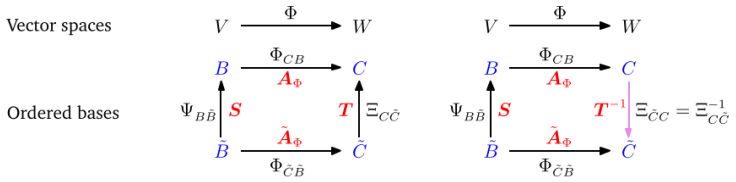
$$\tilde{\boldsymbol{A}}_{\boldsymbol{\Phi}} = \boldsymbol{T}^{-1} \boldsymbol{A}_{\boldsymbol{\Phi}} \boldsymbol{S} \quad \square$$

A.1.11 Intuition for Basis Changes, Equivalence, Similarity

Basis Changes as linear mappings
(A.1.10) tells us that with a basis change in V (B being replaced with $\tilde{B}$) and W (C being replaced with $\tilde{C}$), the transformation matrix $\boldsymbol{A}_\Phi$ of a linear mapping $\Phi : V \to W$ is replaced by an equivalent matrix $\tilde{\boldsymbol{A}}_\Phi$ where

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

Consider a homomorphism $\Phi : V \to W$ and ordered bases $B, \tilde{B}$ of V and $C, \tilde{C}$ of W. The linear mapping Φ_{CB} is an instantiation of Φ and maps basis vectors B onto linear combinations of basis vectors of C. Assuming that we know the transformation matrix $\boldsymbol{A}_\Phi$ of Φ_{CB} with respect to ordered bases B, C, when we perform a basis change from B to $\tilde{B}$ in V and from C to $\tilde{C}$ in W, we can determine the corresponding transformation matrix $\tilde{\boldsymbol{A}}_\Phi$ as follows



First we find the matrix representaion of the linear mapping $\Psi_{B\tilde{B}} : V \to V$ that maps coordinates with respect to the 'new' basis $\tilde{B}$ onto the (unique) coordinates with respect to the 'old' basis B (in V). Then we use the transformation matrix $\boldsymbol{A}_\Phi$ of of $\Phi_{CB} : V \to W$ to map these coordinates onto the coordinates with respect to C in W. Finally we use the inverse of the linear mapping $\Xi_{C\tilde{C}} : W \to W$ that maps the coordinates with respect to $\tilde{C}$ onto the coordinates with respect to C. Thus the linear mapping $\Phi_{\tilde{C}\tilde{B}}$ can be expressed as a composition of linear mappings (that involve the 'old' basis):

$$\Phi_{\tilde{C}\tilde{B}} = \Xi^{-1}_{C\tilde{C}} \circ \Phi_{CB} \circ \Psi_{B\tilde{B}}$$

Concretely, $\Psi_{B\tilde{B}} = \mathrm{id}_V$ and $\Xi_{C\tilde{C}} = \mathrm{id}_W$; they are identity mappings that map vectors onto themselves, but with respect to a different basis. (*coordinate* expression of the vector changes, but the vector itself doesn't change)

Observe that the expression we derived

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

is simply the matrix representation of these mappings.
(next page)

Equivalence and Similarity

- Two matrices $\boldsymbol{A}, \tilde{\boldsymbol{A}} \in \mathbb{R}^{m \times n}$ are *equivalent* if there exist regular/invertible matrices $\boldsymbol{S} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{T} \in \mathbb{R}^{m \times m}$, such that $\tilde{\boldsymbol{A}}_\Phi = \boldsymbol{T}^{-1} \boldsymbol{A}_\Phi \boldsymbol{S}$. (both matrices map vectors from and to the same vector spaces, just using different basis vectors)
- Two matrices $\boldsymbol{A}, \tilde{\boldsymbol{A}} \in \mathbb{R}^{n \times n}$ are *similar* if there exists a regular matrix $\boldsymbol{S} \in \mathbb{R}^{n \times n}$ where $\tilde{\boldsymbol{A}}_\Phi = \boldsymbol{S}^{-1} \boldsymbol{A}_\Phi \boldsymbol{S}$. (the identity mapping for changing bases in the initial and final vector spaces are the same)

Observe that similar matrices are always equivalent, but equivalent matrices are not necessarily similar.

Matrix representation of Basis Changes
Consider vector spaces V, W, X. We know that for linear mappings $\Phi : V \to W$ and $\Psi : W \to X$ the mapping $\Psi \circ \Phi : V \to X$ is also linear. With transformation matrices $\boldsymbol{A}_\Phi$ and $\boldsymbol{A}_\Psi$ of the corresponding mappings, the overall transformation matrix is $\boldsymbol{A}_{\Psi \circ \Phi} = \boldsymbol{A}_\Psi \boldsymbol{A}_\Phi$. Now look at the matrices that represent basis changes from the perspectiveof composing linear mappings:

- $\boldsymbol{A}_\Phi$ is the transformation matrix of a linear mapping $\Phi_{CB} : V \to W$ with respect to the bases B, C.
- $\tilde{\boldsymbol{A}}_\Phi$ is the transformation matrix of the linear mapping $\Phi_{\tilde{C}\tilde{B}} : V \to W$ with respect to the bases $\tilde{B}, \tilde{C}$.
- $\boldsymbol{S}$ is the transformation matrix of a linear mapping $\Psi_{B\tilde{B}} : V \to V$ (automorphism) that represents $\tilde{B}$ in terms of B. Normally, $\Psi = \mathrm{id}_V$ is the identity mapping in V.
- $\boldsymbol{T}$ is the transformation matrix of a linear mapping $\Xi_{C\tilde{C}} : W \to W$ (automorphism) that represents $\tilde{C}$ in terms of C. Normally $\Xi = \mathrm{id}_W$ is the identity mapping in W.

(Informally) writing down the transformations just in terms of bases, one can intuitively see how the derived matrix representation coincdes with each linear mapping:

$$\tilde{B} \to \tilde{C} = \tilde{B} \to B \to C \to \tilde{C}$$
$$\tilde{\boldsymbol{A}}_\Phi = \boldsymbol{T}^{-1} \boldsymbol{A}_\Phi \boldsymbol{S}$$

A.1.12 Image and Kernel

Definition
For $\Phi : V \to W$, we define the *kernel/null space*

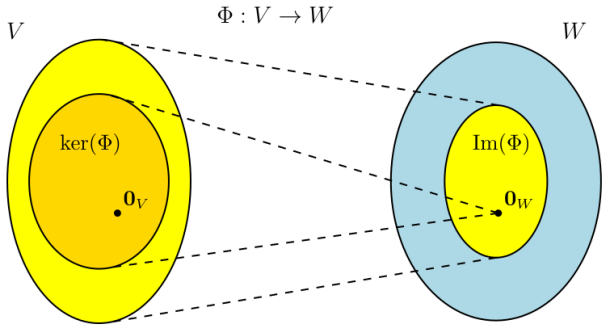
$$\ker(\Phi) := \Phi^{-1}(\mathbf{0}_W) = \{\boldsymbol{v} \in V : \Phi(\boldsymbol{v}) = \mathbf{0}_W\}$$

and the *image/range*

$$\mathrm{Im}(\Phi) := \Phi(V) = \{\boldsymbol{w} \in W | \exists \boldsymbol{v} \in V : \Phi(\boldsymbol{v}) = \boldsymbol{w}\}$$

We also call V and W the *domain* and *codomain* of Φ, respectively.

Intuitively, the kernel is the set of vectors $\boldsymbol{v} \in V$ that Φ maps onto the neutral element $\mathbf{0}_W \in W$. The image is the set of vectors $\boldsymbol{w} \in W$ that can be 'reached' by Φ from any vector in V:



Properties
Consider a linear mapping $\Phi : V \to W$, where V, W are vector spaces.

- It always holds that $\Phi(\mathbf{0}_V) = \mathbf{0}_W$ and therefore $\mathbf{0}_V \in \ker(\Phi)$. This also means the null space is never empty.
- $\mathrm{Im}(\Phi) \subseteq W$ is a subspace of W, and $\ker(\Phi) \subseteq V$ is a subspace of V.
- Φ is injective (one-to-one) if and only if $\ker(\Phi) = \{\mathbf{0}\}$.

(next page)

Null Space and Column Space
Consider $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and a linear mapping $\Phi : \mathbb{R}^n \to \mathbb{R}^m, x \mapsto \boldsymbol{A}x$.

- For $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n]$, where $\boldsymbol{a}_i$ are the columns of $\boldsymbol{A}$, we obtain

$$\begin{aligned}\mathrm{Im}(\Phi) &= \{\boldsymbol{A}\boldsymbol{x} : \boldsymbol{x} \in \mathbb{R}^n\} = \left\{\sum_{i=1}^{n} x_i \boldsymbol{a}_i : x_1, \ldots, x_n \in \mathbb{R}, \boldsymbol{a}_i \in \mathbb{R}^m\right\} \\ &= \mathrm{span}[\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n] \subseteq \mathbb{R}^m\end{aligned}$$

  Essentially this means the image is the span of the columns of $\boldsymbol{A}$, also called the *column space*. Therefore, the column space (image) is a subspace of $\mathbb{R}^m$, where m is the 'height' of the matrix.

- $\mathrm{rk}(A) = \dim(\mathrm{Im}(\Phi))$.
- The kernel/null space $\ker(\Phi)$ is the general solution to the homogeneous system of linear equations $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$; it captures all possible linear combinations of the elements in $\mathbb{R}^n$ that produce $\boldsymbol{0} \in \mathbb{R}^m$.
- The kernel is a subspace of $\mathbb{R}^n$, where n is the 'width' of the matrix.
- The kernel can be used to determine whether/how a column can be expressed as a linear combination of other columns.

Example: Consider

$$\begin{aligned}\Phi : \mathbb{R}^4 \to \mathbb{R}^2, \quad \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} &\mapsto \begin{bmatrix} 1 & 2 & -1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \\ &= x_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + x_2 \begin{bmatrix} 2 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} -1 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} 0 \\ 1 \end{bmatrix}\end{aligned}$$

We get the image from the span of the columns of the transformation matrix, also called the column space.

$$\mathrm{Im}(\Phi) = \mathrm{span}[\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}]$$

To compute the kernel/null space of Φ we solve $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$, this can be done by Gaussian elimination:

$$\begin{bmatrix} 1 & 2 & -1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \rightsquigarrow \cdots \rightsquigarrow \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & -1/2 & -1/2 \end{bmatrix}$$

which gives us (either by observation or with the Minus-1 Trick)

$$\ker(\Phi) = \mathrm{span}[\begin{bmatrix} 0 \\ 1/2 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 1/2 \\ 0 \\ 1 \end{bmatrix}]$$

(next page)

Rank-Nullity Theorem
Theorem: *For vector spaces V, W and a linear mapping $\Phi : V \to W$ it holds that*

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

(also called the *fundamental theorem of linear mappings)* This leads to the following consequences:

- If $\dim(\mathrm{Im}(\Phi)) < \dim(V)$, then $\ker(\Phi)$ is non-trivial, meaning the kernel contains more than $\mathbf{0}_V$ and $\dim(\ker(\Phi)) \geq 1$.
- If $\boldsymbol{A}_\Phi$ is a transformation matrix of Φ with respect to an ordered basis and $\dim(\mathrm{Im}(\Phi)) < \dim(V)$, then the system of linear equations $\boldsymbol{A}_\Phi \boldsymbol{x} = \mathbf{0}$ has infinitely many solutions.
- If $\dim(V) = \dim(W)$, then the three-way equivalence

$$\Phi \text{ is injective} \iff \Phi \text{ is surjective} \iff \Phi \text{ is bijective}$$

holds since $\mathrm{Im}(\Phi) \subseteq W$ (and $\dim(\mathrm{Im}(\Phi)) = \dim(V)$) (see (A.1.8)).

A.1.13 Affine Spaces

Definition
Let V be a vector space, $\boldsymbol{x}_0 \in V$ and $U \subseteq V$ a subspace. Then the subset

$$\begin{aligned} L = \boldsymbol{x}_0 + U &:= \{\boldsymbol{x}_0 + \boldsymbol{u} : \boldsymbol{u} \in U\} \\ &= \{\boldsymbol{v} \in V | \exists \boldsymbol{u} \in U : \boldsymbol{v} = \boldsymbol{x}_0 + \boldsymbol{u}\} \subseteq V \end{aligned}$$

is called an *affine subspace/linear manifold* of V. U is called the *direction or direction space*, and $\boldsymbol{x}_0$ is called the *support point*. (Affine subspaces can also be referred to as *hyperplanes*).

Note that the definition of an affine subspace excludes $\mathbf{0}$ if $\boldsymbol{x}_0 \notin U$. Therefore an affine subspace is not a linear (vector) subspace of V for $\boldsymbol{x}_0 \notin U$. (due to lack of a neutral element)

Subsets of Affine Spaces
Consider two affine spaces $L = \boldsymbol{x}_0 + U$ and $\tilde{L} = \tilde{\boldsymbol{x}}_0 + \tilde{U}$ of a vector space V. Then $L \subseteq \tilde{L}$ if and only if $U \subseteq \tilde{U}$ and $\boldsymbol{x}_0 - \tilde{\boldsymbol{x}}_0 \in \tilde{U}$.

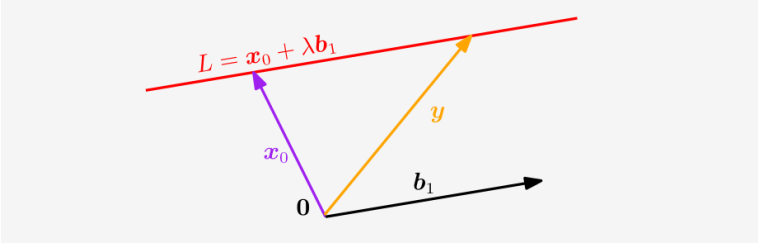
Parametric representation
Affine subspaces are often described by *parameters*: Consider a k-dimensional affine space $L = \boldsymbol{x}_0 + U$ of V. If $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k)$ is an ordered basis of U, then every element $\boldsymbol{x} \in L$ can be uniquely described as

$$\boldsymbol{x} = \boldsymbol{x}_0 + \lambda_1 \boldsymbol{b}_1 + \ldots + \lambda_k \boldsymbol{b}_k$$

where $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$. This representation is called the *parametric equation* of L with directional vectors $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k$ and *parameters* $\lambda_1, \ldots, \lambda_k$.
(next page)

Examples of Affine Subspaces

- One-dimensional affine subspaces are called *lines*, and can be written parametrically as $\boldsymbol{y} = \boldsymbol{x}_0 + \lambda_1 \boldsymbol{b}_1$, where $\lambda \in \mathbb{R}$ and $U = \text{span}[\boldsymbol{b}_1] \subseteq \mathbb{R}^n$ is a one dimensional subspace of $\mathbb{R}^n$. Illustrated:



- Two-dimensional affine subspaces of $\mathbb{R}^n$ are called *planes*, with parametric equation $\boldsymbol{y} = \boldsymbol{x}_0 + \lambda_1 \boldsymbol{b}_1 + \lambda_2 \boldsymbol{b}_2$, where $\lambda_1, \lambda_2 \in \mathbb{R}$ and $U = \text{span}[\boldsymbol{b}_1, \boldsymbol{b}_2] \subseteq \mathbb{R}^n$.

- In $\mathbb{R}^n$, the $(n-1)$-dimensional affine subspaces are called *hyperplanes*, with parametric equation $\boldsymbol{y} = \boldsymbol{x}_0 + \sum_{i=1}^{n-1} \lambda_i \boldsymbol{b}_i$, where $\boldsymbol{b}_1, \ldots \boldsymbol{b}_{n-1}$ form a basis of an $(n-1)$-dimensional subspace U of $\mathbb{R}^n$. (in $\mathbb{R}^2$, a line is a hyperplane; in $\mathbb{R}^3$, a plane is a hyperplane.

Inhomogeneous systems of linear equations and affine subspaces
For $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{x} \in \mathbb{R}^m$, the solution of the system of linear equations $\boldsymbol{A\lambda} = \boldsymbol{x}$ is either the empty set or an affine subspace of $\mathbb{R}^n$ of dimension $n - \text{rk}(A)$; the solution of the linear equation $\lambda_1 \boldsymbol{b}_1 + \ldots + \lambda_n \boldsymbol{b}_n = \boldsymbol{x}$, where $(\lambda_1, \ldots, \lambda_n) \neq (0, \ldots, 0)$, is a hyperplane in $\mathbb{R}^n$. (I assume solutions for one-to-one mappings are also affine spaces with direction $\{\mathbf{0}\}$ and support vector equal to the solution).

Viewed in another way, every k-dimensional affine subspace is the solution to an inhomogeneous system $\boldsymbol{Ax} = \boldsymbol{b}$, where $\boldsymbol{A} \in \mathbb{R}^{m \times n}, \boldsymbol{b} \in \mathbb{R}^m$ and $\text{rk}(A) = n - k$. (The solution to $\boldsymbol{Ax} = \mathbf{0}$ can also be thought of as a special affine subspace with support point $\boldsymbol{x}_0 = \mathbf{0}$.)
(next page)

Affine Mappings
Definition
For two vector spaces V, W, a linear mapping $\Phi : V \to W$, and $\boldsymbol{a} \in W$, the mapping

$$\begin{aligned} \phi :& V \to W \\ & \boldsymbol{x} \mapsto \boldsymbol{a} + \Phi(\boldsymbol{x}) \end{aligned}$$

is an *affine mapping* from V to W. The vector $\boldsymbol{a}$ is called the *translation vector* of ϕ. Note that

- Every affine mapping $\phi : V \to W$ is also a composition of linear mapping $\Phi : V \to W$ and a translation $\tau : W \to W$ in W, such that $\phi = \tau \circ \Phi$. The mappings Φ and τ are uniquely determined.
- The composition $\phi' \circ \phi$ of affine mappings $\phi : V \to W$, $\phi' : W \to X$ is affine.
- If ϕ is bijective, affine mappings keep the geometric structure invariant. They also preserve the dimension and parallelism.

A.2 Analytic Geometry

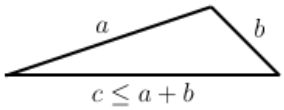
A.2.1 Norms

Definition
A *norm* on a vector space V is a function

$$\begin{aligned} ||\cdot|| :& V \to \mathbb{R} \\ & \boldsymbol{x} \mapsto ||\boldsymbol{x}|| \end{aligned}$$

which assigns each vector $\boldsymbol{x}$ its *length* $||\boldsymbol{x}|| \in \mathbb{R}$, such that for all $\lambda \in \mathbb{R}$ and $\boldsymbol{x}, \boldsymbol{y} \in V$ the following hold:

- *Absolutely homogeneous*: $||\lambda \boldsymbol{x}|| = |\lambda| \, ||\boldsymbol{x}||$
- *Triangle inequality*: $||\boldsymbol{x} + \boldsymbol{y}|| \leq ||\boldsymbol{x}|| + ||\boldsymbol{y}||$



- *Positive definite*: $||\boldsymbol{x}|| \geq 0$ and $||\boldsymbol{x}|| = 0 \implies \boldsymbol{x} = \boldsymbol{0}$

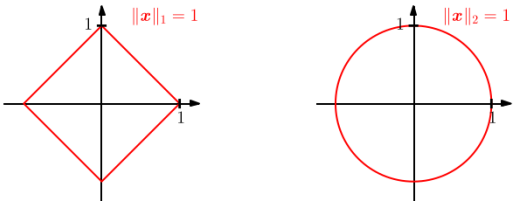
Example: Manhattan and Euclidean Norm
The *Manhattan Norm* on $\mathbb{R}^n$ is defined for $\boldsymbol{x} \in \mathbb{R}^n$ as

$$||\boldsymbol{x}||_1 := \sum_{i=1}^{n} |x_i|$$

where $|\cdot|$ is the absolute value. The *Euclidean norm* for $\boldsymbol{x} \in \mathbb{R}^n$ is defined as

$$||\boldsymbol{x}||_2 := \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{\boldsymbol{x}^T \boldsymbol{x}}$$

giving us the *Euclidean distance* of $\boldsymbol{x}$ from the origin. The Euclidean norm is also called the ℓ_2 norm.



Vectors $\boldsymbol{x} \in \mathbb{R}^2$ with $||\boldsymbol{x}||_1 = 1$ (left) and $||\boldsymbol{x}||_2 = 1$ (right)

Generally the Euclidean norm is used by default if not stated otherwise.

A.2.2 Inner Products

Dot Product
The *scalar/dot product* in $\mathbb{R}^n$ is given by

$$\boldsymbol{x}^T \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i$$

The Dot product is a type of inner product, but inner products are actually more general concepts with specific properties.

General Inner Products
Linear mappings can be rearranged with respect to addition and multiplication with a scalar. A *bilinear mapping* Ω is a mapping with two arguments, and is linear in each argument; for instance considering a vector space V, it holds that for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in V$, $\lambda, \psi \in \mathbb{R}$ that

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

point here is that Ω is linear in both arguments.

Definition: Symmetric, Positive definite
Let V be a vector space and $\Omega : V \times V \to \mathbb{R}$ be a bilinear mapping that takes two vectors and maps them onto a real number. Then

- Ω is called *symmetric* if $\Omega(\boldsymbol{x}, \boldsymbol{y}) = \Omega(\boldsymbol{y}, \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y} \in V$; essentially the order of the arguments does not matter.
- Ω is called *positive definite* if

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

Definition: Inner Product, Inner Product Space
Let V be a vector space and $\Omega : V \times V \to \mathbb{R}$ be a bilinear mapping that takes two vectors and maps them onto a real number. Then

- A positive definite, symmetric bilinear mapping $\Omega : V \times V \to \mathbb{R}$ is called an *inner product* on V, typically written $\langle x, y \rangle$ instead of $\Omega(x, y)$.
- The pair $(V, \langle \cdot, \cdot \rangle)$ is called an *inner product space* or (real) *vector space with inner product*. If the inner product is the dot product, we call $(V, \langle \cdot, \cdot \rangle)$ a *Euclidean vector space*.

A.2.3 Symmetric, Positive Definite Matrices

Consider an n-dimensional vector space V with an inner product $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ and an ordered basis $B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)$ of V. Since any vectors $\boldsymbol{x}, \boldsymbol{y} \in V$ can be written as linear combinations of the basis vectors such that

$$\boldsymbol{x} = \sum_{i=1}^{n} \psi_i \boldsymbol{b}_i \in V, \quad \text{and} \quad \boldsymbol{y} = \sum_{j=1}^{n} \lambda_j \boldsymbol{b}_j \in V$$

for suitable $\psi_i, \lambda_j \in \mathbb{R}$, due to the bilinearity of the inner product it holds that for all $\boldsymbol{x}, \boldsymbol{y} \in V$ that

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \left\langle \sum_{i=1}^{n} \psi_i \boldsymbol{b}_i, \sum_{j=1}^{n} \lambda_j \boldsymbol{b}_j \right\rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_i \langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle \lambda_j = \hat{\boldsymbol{x}}^T \boldsymbol{A} \hat{\boldsymbol{y}}$$

where $A_{ij} := \langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle$ and $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ are the coordinates of $\boldsymbol{x}$ and $\boldsymbol{y}$ with respect to the basis B.

This implies that the inner product $\langle \cdot, \cdot \rangle$ is uniquely determined through $\boldsymbol{A}$. The symmetry of the inner product also means that $\boldsymbol{A}$ is symmetric. Furthermore, the positive definiteness of the inner product implies that

$$\forall \boldsymbol{x} \in V \backslash \{\boldsymbol{0}\} : \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} > 0$$

Definition
Following from the above reasoning, we define a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ that satisfies

$$\forall \boldsymbol{x} \in V \backslash \{\boldsymbol{0}\} : \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} > 0$$

to be *symmetric, positive definite* or just *positive definite*. If only $\geq$ holds then $\boldsymbol{A}$ is called *symmetric, positive semidefinite*.

Example: Considering the matrices

$$\boldsymbol{A}_1 = \begin{bmatrix} 9 & 6 \\ 6 & 5 \end{bmatrix}, \quad \begin{bmatrix} 9 & 6 \\ 6 & 3 \end{bmatrix}$$

$\boldsymbol{A}_1$ is positive definite because it is symmetric and for all $\boldsymbol{x} \in V \backslash \{\boldsymbol{0}\}$

$$\begin{aligned} \boldsymbol{x}^T \boldsymbol{A}_1 \boldsymbol{x} &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 9 & 6 \\ 6 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= 9x_1^2 + 12x_1x_2 + 5x_2^2 = (3x_1 + 2x_2)^2 + x_2^2) > 0 \end{aligned}$$

While $\boldsymbol{A}_2$ is symmetric but not positive definite because

$$\boldsymbol{x}^T \boldsymbol{A}_2 \boldsymbol{x} = 9x_1^2 + 12x_1x_2 + 3x_2^2 = (3x_1 + 2x_2)^2 - x_2^2)$$

which isn't always greater than 0 for all $\boldsymbol{x} \in V \backslash \{\boldsymbol{0}\}$.
(next page)

Significance of symmetric positive definite matrices
If $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is symmetric, positive definite, then

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \hat{\boldsymbol{x}}^T \boldsymbol{A} \tilde{\boldsymbol{y}}$$

defines an inner product with respect to an ordered basis B, where $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ are the coordinate representations of $\boldsymbol{x}, \boldsymbol{y} \in V$ with respect to B.

Theorem 3.5: *For a real-valued, finite-dimensional vector space V and an ordered basis B of V, it holds that $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ is an inner product if and only if there exists a symmetric, positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ with*

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \hat{\boldsymbol{x}}^T \boldsymbol{A} \tilde{\boldsymbol{y}}$$

If $A \in \mathbb{R}^{n \times n}$ is symmetric and positive definite, the following properties hold:

- The null space/kernel of $\boldsymbol{A}$ consists only of $\boldsymbol{0}$ because $\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} > 0$ for all $\boldsymbol{x} \neq \boldsymbol{0}$. This implies that $\boldsymbol{A}\boldsymbol{x} \neq \boldsymbol{0}$ if $\boldsymbol{x} \neq \boldsymbol{0}$.
- The diagonal elements a_{ii} of $\boldsymbol{A}$ are positive because $a_{ii} = \boldsymbol{e}_i^T \boldsymbol{A} \boldsymbol{e}_i > 0$, where $\boldsymbol{e}_i$ is the i-th vector of the standard basis in $\mathbb{R}^n$. (Alternatively, see that $\boldsymbol{A}$ is just made up of the inner product of different combinations of the basis vectors, where the diagonal only has inner products between each basis vector and itself; the positive definite requirement of the inner product therefore means that the diagonal will be positive.)

A.2.4 Cauchy-Schwarz Inequality

For an inner product vector space $(V, \langle \cdot, \cdot \rangle)$ the induced norm satisfies the *Cauchy-Schwarz inequality*

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

The equality only holds if $\boldsymbol{x}$ and $\boldsymbol{y}$ are linearly dependent.

Proof
Let V be a vector space over the real or complex field F, and let $\boldsymbol{x}, \boldsymbol{y} \in V$. First we show that $|\langle \boldsymbol{x}, \boldsymbol{y} \rangle|^2 = \langle \boldsymbol{x}, \boldsymbol{x} \rangle \langle \boldsymbol{y}, \boldsymbol{y} \rangle$ if $\boldsymbol{y} = a\boldsymbol{x}$ for some $a \in F$ (linearly dependent). Then we show $|\langle \boldsymbol{x}, \boldsymbol{y} \rangle|^2 < \langle \boldsymbol{x}, \boldsymbol{x} \rangle \langle \boldsymbol{y}, \boldsymbol{y} \rangle$ if $\boldsymbol{y} \neq a\boldsymbol{x}$ for all $a \in F$ (linearly independent).

If $\boldsymbol{y} = a\boldsymbol{x}$ for some $a \in F$

$$\begin{aligned} |\langle \boldsymbol{x}, \boldsymbol{y} \rangle|^2 &= |\langle \boldsymbol{x}, a\boldsymbol{x} \rangle|^2 \\ &= |a \langle \boldsymbol{x}, \boldsymbol{x} \rangle|^2 \quad \text{(linearity)} \\ &= |a|^2 \langle \boldsymbol{x}, \boldsymbol{x} \rangle^2 \end{aligned}$$

similarly it follows that

$$\begin{aligned} \langle \boldsymbol{x}, \boldsymbol{x} \rangle \langle \boldsymbol{y}, \boldsymbol{y} \rangle &= \langle \boldsymbol{x}, \boldsymbol{x} \rangle \langle a\boldsymbol{x}, a\boldsymbol{x} \rangle \\ &= \langle \boldsymbol{x}, \boldsymbol{x} \rangle \, a^2 \langle \boldsymbol{x}, \boldsymbol{x} \rangle \\ &= |a|^2 \langle \boldsymbol{x}, \boldsymbol{x} \rangle^2 \end{aligned}$$

giving us $|\langle \boldsymbol{x}, \boldsymbol{y} \rangle|^2 = \langle \boldsymbol{x}, \boldsymbol{x} \rangle \langle \boldsymbol{y}, \boldsymbol{y} \rangle$ if $\boldsymbol{y} = a\boldsymbol{x}$ for some $a \in F$ (linearly dependent). Now considering the case where $\boldsymbol{y} \neq a\boldsymbol{x}$ for all $a \in F$; it is implied that $\boldsymbol{y} \neq 0$ (because of a) and that therefore $\langle \boldsymbol{y}, \boldsymbol{y} \rangle \neq 0$. We can also say that for all $a \in F$, $\langle \boldsymbol{x} - a\boldsymbol{y}, \boldsymbol{x} - a\boldsymbol{y} \rangle > 0$. Now we expand:

$$\begin{aligned} \langle \boldsymbol{x} - a\boldsymbol{y}, \boldsymbol{x} - a\boldsymbol{y} \rangle &= \langle \boldsymbol{x}, \boldsymbol{x} - a\boldsymbol{y} \rangle - a \langle \boldsymbol{y}, \boldsymbol{x} - a\boldsymbol{y} \rangle \\ &= \langle \boldsymbol{x}, \boldsymbol{x} \rangle - a \langle \boldsymbol{x}, \boldsymbol{y} \rangle - a \langle \boldsymbol{y}, \boldsymbol{x} \rangle + a^2 \langle \boldsymbol{y}, \boldsymbol{y} \rangle \end{aligned}$$

Choosing $a = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\langle \boldsymbol{y}, \boldsymbol{y} \rangle}$:

$$\begin{aligned} \langle \boldsymbol{x} - a\boldsymbol{y}, \boldsymbol{x} - a\boldsymbol{y} \rangle &= \langle \boldsymbol{x}, \boldsymbol{x} \rangle - \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\langle \boldsymbol{y}, \boldsymbol{y} \rangle} \langle \boldsymbol{x}, \boldsymbol{y} \rangle - \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\langle \boldsymbol{y}, \boldsymbol{y} \rangle} \langle \boldsymbol{y}, \boldsymbol{x} \rangle + \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle^2}{\langle \boldsymbol{y}, \boldsymbol{y} \rangle^2} \langle \boldsymbol{y}, \boldsymbol{y} \rangle \\ &= \langle \boldsymbol{x}, \boldsymbol{x} \rangle - \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle \langle \boldsymbol{y}, \boldsymbol{x} \rangle}{\langle \boldsymbol{y}, \boldsymbol{y} \rangle} \end{aligned}$$

(next page)

Since we have $\langle \boldsymbol{x} - a\boldsymbol{y}, \boldsymbol{x} - a\boldsymbol{y} \rangle > 0$ and

$$\langle \boldsymbol{x} - a\boldsymbol{y}, \boldsymbol{x} - a\boldsymbol{y} \rangle = \langle \boldsymbol{x}, \boldsymbol{x} \rangle - \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle \langle \boldsymbol{y}, \boldsymbol{x} \rangle}{\langle \boldsymbol{y}, \boldsymbol{y} \rangle}$$

we have the inequality

$$\begin{aligned} \langle \boldsymbol{x}, \boldsymbol{x} \rangle - \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle \langle \boldsymbol{y}, \boldsymbol{x} \rangle}{\langle \boldsymbol{y}, \boldsymbol{y} \rangle} &> 0 \\ \langle \boldsymbol{x}, \boldsymbol{x} \rangle \langle \boldsymbol{y}, \boldsymbol{y} \rangle - \langle \boldsymbol{x}, \boldsymbol{y} \rangle \langle \boldsymbol{y}, \boldsymbol{x} \rangle &> 0 \\ |\langle \boldsymbol{x}, \boldsymbol{y} \rangle|^2 &< \langle \boldsymbol{x}, \boldsymbol{x} \rangle \langle \boldsymbol{y}, \boldsymbol{y} \rangle \end{aligned}$$

where $\boldsymbol{y} \neq a\boldsymbol{x}$ for all $a \in F$.

Therefore the inequality

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle|^2 \leq \langle \boldsymbol{x}, \boldsymbol{x} \rangle \langle \boldsymbol{y}, \boldsymbol{y} \rangle$$

holds for all $\boldsymbol{x}, \boldsymbol{y} \in V$, with the equality holding only if they are linearly dependent.

Since the norm is induced by the inner product

$$||\boldsymbol{x}|| := \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$$

we can also say

$$\begin{aligned} |\langle \boldsymbol{x}, \boldsymbol{y} \rangle|^2 &\leq ||\boldsymbol{x}||^2 ||\boldsymbol{y}||^2 \\ |\langle \boldsymbol{x}, \boldsymbol{y} \rangle| &\leq ||\boldsymbol{x}||\,||\boldsymbol{y}|| \end{aligned}$$

A.2.5 Lengths and Distances

Inner products and norms are closely related in the sense that any inner product induces a norm, which represent the length of a vector:

$$||\boldsymbol{x}|| := \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$$

Different inner products therefore lead to different norms—different representations of the length of a vector. Note however that not every norm is induced by an inner product (take the Manhattan norm for instance).

Distance and Metric
Considering an inner product space $(V, \langle \cdot, \cdot \rangle)$,

$$d(\boldsymbol{x}, \boldsymbol{y}) := ||\boldsymbol{x} - \boldsymbol{y}|| = \sqrt{\langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{y} \rangle}$$

is called the *distance* between $\boldsymbol{x}$ and $\boldsymbol{y}$ for $\boldsymbol{x}, \boldsymbol{y} \in V$. If the dot product is used as the inner product, then the distance is called *Euclidean distance*.

The mapping

$$\begin{aligned} d : V \times V &\to \mathbb{R} \\ (\boldsymbol{x}, \boldsymbol{y}) &\mapsto d(\boldsymbol{x}, \boldsymbol{y}) \end{aligned}$$

is called a metric. (remember that the distance between vectors does not require an inner product; a norm is sufficient, but that an inner product, when defined, would induce a norm).

A metric d satisfies the following:

1. d is *positive definite*, meaning $d(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ for all $\boldsymbol{x}, \boldsymbol{y} \in V$ and $d(\boldsymbol{x}, \boldsymbol{y}) = 0 \iff \boldsymbol{x} = \boldsymbol{y}$.
2. d is *symmetric*, meaning $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y} \in V$.
3. *Triangle inequality*: $d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$ for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in V$.

At first glance the properties of inner products and metrics appear similar. Note however that the properties of metrics come from those of norms rather than inner products (see (A.2.1)). Also notice that similar inputs will result in a small metric but not inner product.

A.2.6 Angles and Orthogonality

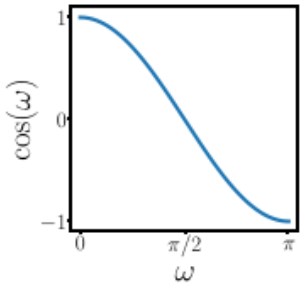
Angle

Inner products also capture the geometry of a vector space by defining the angle ω between two vectors. Using the Cauchy-Schwarz inequality, assuming that $\boldsymbol{x} \neq \mathbf{0}, \boldsymbol{y} \neq \mathbf{0}$,

$$\begin{aligned} |\langle \boldsymbol{x}, \boldsymbol{y} \rangle| &\leq ||\boldsymbol{x}|| \, ||\boldsymbol{y}|| \\ \frac{|\langle \boldsymbol{x}, \boldsymbol{y} \rangle|}{||\boldsymbol{x}|| \, ||\boldsymbol{y}||} &\leq 1 \\ -1 \leq \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{||\boldsymbol{x}|| \, ||\boldsymbol{y}||} &\leq 1 \end{aligned}$$

Thus there exists a unique $\omega \in [0, \pi]$, with

$$\cos \omega = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{||\boldsymbol{x}|| \, ||\boldsymbol{y}||}$$



Each value in $[-1, 1]$ corresponds to a unique ω; the number ω is the *angle* between the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. Intuitively, the angle between two vectors tells us how similar their orientations are. (For instance, using the dot product, the angle between $\boldsymbol{x}$ and $\boldsymbol{y} = 4\boldsymbol{x}$ is 0.)

(next page)

Orthogonality
The *angle* ω in

$$\cos\omega = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{||\boldsymbol{x}||\,||\boldsymbol{y}||}$$

can be intuitively seen as a measure of how similar the orientation between two vectors are. (The angle between a vector and itself is 0, for instance.)

Definition
Two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are *orthogonal* if and only if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$, written as $\boldsymbol{x} \perp \boldsymbol{y}$. If additionally $||\boldsymbol{x}|| = 1 = ||\boldsymbol{y}||$ (unit vectors), then $\boldsymbol{x}$ and $\boldsymbol{y}$ are *orthonormal*. (Notice that the $\mathbf{0}$-vector is orthogonal to every vector in the vector space.)

Orthogonality is the generalisation of the concept of perpendicularity to bilinear forms that do not have to be the dot product. Geometrically, one can think of orthogonal vectors as having a right angle with respect to a *specific* inner product.

Example
Consider two vectors $\boldsymbol{x} = [1, 1]^T$,$\boldsymbol{y} = [-1, 1]^T \in \mathbb{R}^2$. We are interested in determining the angle ω between them using two different inner products. Using the dot product as the inner product yields an angle ω between $\boldsymbol{x}$ and $\boldsymbol{y}$ of $90°$; in this case $\boldsymbol{x} \perp \boldsymbol{y}$. However, considering a different inner product

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \boldsymbol{y}$$

we get a different angle ω:

$$\cos\omega = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{||\boldsymbol{x}||\,||\boldsymbol{y}||} = -\frac{1}{3} \implies \omega \approx 1.91\,\text{rad} \approx 109.5°$$

A.2.7 Orthogonal Matrices

Definition
A square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is an *orthogonal matrix* if and only if its columns are **orthonormal** so that

$$\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I} = \boldsymbol{A}^T\boldsymbol{A}$$

which implies that

$$\boldsymbol{A}^{-1} = \boldsymbol{A}^T$$

meaning the inverse can be obtained by transposing the matrix.

Side note (250724):
'Orthogonal matrices' suggests some general inner product expansion involved in matrix multiplication. But notice that orthogonality between columns only holds for the standard inner product (the dot product); for intuition, consider an inner product

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T\boldsymbol{M}\boldsymbol{y}$$

we only have $\boldsymbol{A} = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)$ being orthonormal if

$$\forall i, j : \langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle = \boldsymbol{e}_i^T\boldsymbol{M}\boldsymbol{e}_j = \delta_{i,j}$$

where $\delta_{i,j}$ is the kronecker delta. Written compactly in matrix notation:

$$\boldsymbol{A}^T\boldsymbol{M}\boldsymbol{A} = \boldsymbol{I}$$

only for the standard inner product $\boldsymbol{M} = \boldsymbol{I}$ do we obtain $\boldsymbol{A}^T\boldsymbol{A}$.

(A better understanding comes from a definition for orthogonality of linear transformations: *On a finite-dimensional inner product space* $(\mathbb{V}, \langle \cdot, \cdot \rangle)$*, a linear transformation* $T : \mathbb{V} \to \mathbb{V}$ *is said to be* ***orthogonal*** *if it preserves the inner product, that is, if* $\langle T(x), T(y) \rangle = \langle \boldsymbol{x}, \boldsymbol{y} \rangle$; the definition $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I} = \boldsymbol{A}^T\boldsymbol{A}$ comes from $\boldsymbol{A}$ being a matrix representation of such a mapping (thus the added requirement of orthonormality)).
(next page)

Properties

Transformations by orthogonal matrices are special because the length of a vector x is not changed after the transformation. For the dot product we obtain (notice that orthogonality of A depends on the choice of inner product)

$$||\boldsymbol{Ax}||^2 = (\boldsymbol{Ax})^T(\boldsymbol{Ax}) = \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{Ax} = \boldsymbol{x}^T\boldsymbol{Ix} = \boldsymbol{x}^T\boldsymbol{x} = ||\boldsymbol{x}||^2$$

The angle, as measured by the inner product, is also unchanged upon transformation by an orthogonal matrix; assuming the dot product as the inner product:

$$\cos\omega = \frac{(\boldsymbol{Ax}^T)(\boldsymbol{Ay})}{||\boldsymbol{Ax}||\,||\boldsymbol{Ay}||} = \frac{\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{Ay}}{\sqrt{\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{Axy}^T\boldsymbol{A}^T\boldsymbol{Ay}}} = \frac{\boldsymbol{x}^T\boldsymbol{y}}{||\boldsymbol{x}||\,||\boldsymbol{y}||}$$

Orthogonal matrices therefore preserve both angles and distances.

A.2.8 Orthonormal Basis and Complement

Orthogonal Basis
Considering an n-dimensional vector space V and a basis $\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n\}$ of V. If

$$\begin{aligned} \langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle &= 0 \quad (\text{for } i \neq j) \\ \langle \boldsymbol{b}_i, \boldsymbol{b}_i \rangle &= 1 \end{aligned}$$

for all $i, j = 1, \ldots, n$ then the basis is called an *orthonormal basis (ONB)*. If only $\langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle = 0$ for $i \neq j$, then the basis is called an *orthogonal basis*.

Orthogonal Complement
Consider a D-dimensional vector space V and an M-dimenstional subspace $U \in V$. Then its *orthogonal complement* $U^\perp$ is a $(D-M)$-dimensional subspace of V and contains all vectors in V that are orthogonal to every vector in U.

Since $U \cap U^\perp = \{\boldsymbol{0}\}$ any vector $\boldsymbol{x} \in V$ can be uniquely decomposed into

$$\boldsymbol{x} = \sum_{m=1}^{M} \lambda_m \boldsymbol{b}_m + \sum_{j=1}^{D-M} \psi_j \boldsymbol{b}_j^\perp, \quad \lambda_m, \psi_j \in \mathbb{R}$$

where $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M)$ is basis of U and $(\boldsymbol{b}_1^\perp, \ldots, \boldsymbol{b}_{D-M}^\perp)$ is a basis of $U^\perp$.

Orthogonal complements can be used to describe hyperplanes in n-dimensional vector/affine spaces. For instance consider a three dimensional vector space; a plane U in this space can be described by a vector w orthogonal to the it (called the *normal* vector of U). The vector w with $||w|| = 1$ is the basis vector of U^T.

A.2.9 Inner Product of Functions

We can think of a vector $\boldsymbol{x} \in \mathbb{R}^n$ as a function with n function values (outputs). The concept of an inner product can then be generalised to vectors with an infinite number of entries (countably infinite) and also continuous-valued functions (uncountably infinite). The sum over individual components of vector can be expressed as an integral.

An inner product of two functions $u : \mathbb{R} \to \mathbb{R}$ and $v : \mathbb{R} \to \mathbb{R}$ can be defined as the definite integral

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

for lower and upper limits $a, b < \infty$. As with usual inner products, we can define norms and orthogonality from the inner product. (more mathematically precise definitions require real and functional analysis)

Example
If we choose $u = \sin(x)$ and $v = \cos(x)$, the integrand $f(x) = u(x)v(x)$ is odd. Therefore the integral over the limits $a = -\pi$, $b = \pi$ of f evaluates to 0—sin and cos are orthogonal functions.

Also notice that the collection of functions

$$\{1, \cos(x), \cos(2x), \cos(3x), \ldots\}$$

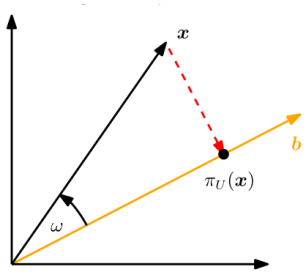
is orthogonal if we integrate over $-\pi$ to π meaning any pair of functions are orthogonal. Projection of functions onto this subspace is the fundamental idea behind Fourier series.

A.2.10 Orthogonal Projections I

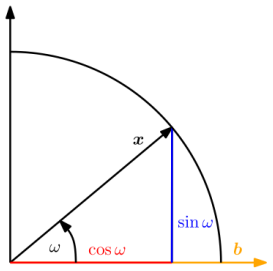
Projection
Let V be a vector space and $U \subseteq V$ a subspace of V. A linear mapping $\pi : V \to U$ is called a *projection* if $\pi^2 = \pi \circ \pi = \pi$.

Linear mappings can be expressed as matrices; correspondingly, projections can be expressed as *projection matrices* $\boldsymbol{P}_\pi$, which exhibit the property $\boldsymbol{P}_\pi^2 = \boldsymbol{P}_\pi$.

Orthogonal projection onto one-dimensional subspaces
Consider a one-dimensional subspace (a line) through the origin with basis vector $\boldsymbol{b} \in \mathbb{R}^n$. Say the line is a one-dimensional subspace $U \subseteq \mathbb{R}^n$ spanned by $\boldsymbol{b}$. When we project $\boldsymbol{x} \in \mathbb{R}^n$ onto U, we are seeking the vector $\pi_U(\boldsymbol{x}) \in U$ that is closest to $\boldsymbol{x}$.



(a) Projection of $\boldsymbol{x} \in \mathbb{R}^2$ onto a subspace U with basis vector $\boldsymbol{b}$.



(b) Projection of a two-dimensional vector $\boldsymbol{x}$ with $\|\boldsymbol{x}\| = 1$ onto a one-dimensional subspace spanned by $\boldsymbol{b}$.

We can characterise the properties of the projection $\pi_U(\boldsymbol{x})$ as the following:

- The projection $\pi_U(\boldsymbol{x})$ is closest to $\boldsymbol{x}$, where 'closest' implies the distance $||\boldsymbol{x} - \pi_U(\boldsymbol{x})||$ is minimal. It follows that the segment $\pi_U(\boldsymbol{x}) - \boldsymbol{x}$ from $\pi_U(\boldsymbol{x})$ to $\boldsymbol{x}$ is orthogonal to U and therefore the basis vector $\boldsymbol{b}$ of U. The orthogonality condition yields $\langle \pi_U(\boldsymbol{x}) - \boldsymbol{x}, \boldsymbol{b} \rangle = 0$ since angles between vectors are defined by the inner product.
- The projection $\pi_U(\boldsymbol{x})$ of $\boldsymbol{x}$ onto U must be an element of U and therefore a multiple of the basis vector $\boldsymbol{b}$ that spans U. Hence $\pi_U(\boldsymbol{x}) = \lambda \boldsymbol{b}$ for some $\lambda \in \mathbb{R}$.

(next page)

Deriving the projection

1. **Determining the coordinate of projection λ**
   The orthogonality condition yields

   $$\langle \boldsymbol{x} - \pi_U(\boldsymbol{x}), \boldsymbol{b} \rangle = 0 \iff \langle \boldsymbol{x} - \lambda \boldsymbol{b}, \boldsymbol{b} \rangle = 0$$

   Exploiting the bilinearity of the inner product:

   $$\langle \boldsymbol{x}, \boldsymbol{b} \rangle - \lambda \langle \boldsymbol{b}, \boldsymbol{b} \rangle = 0 \iff \lambda = \frac{\langle \boldsymbol{x}, \boldsymbol{b} \rangle}{\langle \boldsymbol{b}, \boldsymbol{b} \rangle} = \frac{\langle \boldsymbol{b}, \boldsymbol{x} \rangle}{||\boldsymbol{b}||^2}$$

   If we choose $\langle \cdot, \cdot \rangle$ to be the dot product,

   $$\lambda = \frac{\boldsymbol{b}^T \boldsymbol{x}}{\boldsymbol{b}^T \boldsymbol{b}} = \frac{\boldsymbol{b}^T \boldsymbol{x}}{||\boldsymbol{b}||^2}$$

   if $||\boldsymbol{b}|| = 1$, then the coordinate λ of the projection is given by $\boldsymbol{b}^T \boldsymbol{x}$.

2. **Finding the projection point $\pi_U(\boldsymbol{x})$**
   Since $\pi_U(\boldsymbol{x}) = \lambda \boldsymbol{b}$, we have

   $$\pi_U(\boldsymbol{x}) = \lambda \boldsymbol{b} = \frac{\langle \boldsymbol{b}, \boldsymbol{x} \rangle}{||\boldsymbol{b}||^2} \boldsymbol{b} \underbrace{= \frac{\boldsymbol{b}^T \boldsymbol{x}}{||\boldsymbol{b}||^2} \boldsymbol{b}}_{\text{for the dot product}}$$

   we can also compute the length of $\pi_U(\boldsymbol{x})$:

   $$||\pi_U(\boldsymbol{x})|| = ||\lambda \boldsymbol{b}|| = |\lambda| \, ||\boldsymbol{b}||$$

   our projection is of length $|\lambda|$ times the length of $\boldsymbol{b}$. (This coincides with the idea that λ is the coordinate of $\pi_U(\boldsymbol{x})$ with respect to the basis vector $\boldsymbol{b}$ spanning our one-dimensional subspace U.)

   Using the dot product as an inner product:

   $$||\pi_U(\boldsymbol{x})|| = \frac{|\boldsymbol{b}^T \boldsymbol{x}|}{||\boldsymbol{b}||^2} ||\boldsymbol{b}|| = |\cos \omega| \, ||\boldsymbol{x}||$$

   where ω is the angle between $\boldsymbol{x}$ and $\boldsymbol{b}$; notice this equation intuitively makes sense.
   (next page)

3. **Finding the projection matrix $\boldsymbol{P}_\pi$**
   Since a projection is a linear mapping there exists a projection matrix $\boldsymbol{P}_\pi$ such that $\pi_U(\boldsymbol{x}) = \boldsymbol{P}_\pi \boldsymbol{x}$. For the dot product as inner product:

$$\pi_U(\boldsymbol{x}) = \lambda \boldsymbol{b} = \boldsymbol{b}\lambda = \boldsymbol{b}\frac{\boldsymbol{b}^T\boldsymbol{x}}{||\boldsymbol{b}||^2} = \frac{\boldsymbol{b}\boldsymbol{b}^T}{||\boldsymbol{b}||^2}\boldsymbol{x}$$

   giving us

$$\boldsymbol{P}_\pi = \frac{\boldsymbol{b}\boldsymbol{b}^T}{||\boldsymbol{b}||^2}$$

   Note that $\boldsymbol{b}\boldsymbol{b}^T$ (and consequently $\boldsymbol{P}_\pi$) is a symmetric matrix with rank 1, and $||\boldsymbol{b}||^2 = \langle \boldsymbol{b}, \boldsymbol{b} \rangle$ is a scalar.

The projection matrix $\boldsymbol{P}_\pi$ projects any vector $\boldsymbol{x} \in \mathbb{R}^n$ onto the line through the origin with direction $\boldsymbol{b}$—the subspace U spanned by $\boldsymbol{b}$. Note that the projection $\pi_U(\boldsymbol{x}) \in \mathbb{R}^n$ is still an n-dimensional vector and not a scalar; however we dont require n coordinates to represent the projection—just a single coordinate λ representing it with respect to the basis vector $\boldsymbol{b}$ spanning U.

Example (assuming the dot product as inner product)
Consider finding the projection matrix $\boldsymbol{P}_\pi$ onto the line through the origin spanned by $\boldsymbol{b} = [1, 2, 2]^T$ (meaning $\boldsymbol{b}$ is a direction and a basis of the one-dimensional subspace/line through origin). We have

$$\boldsymbol{P}_\pi = \frac{\boldsymbol{b}\boldsymbol{b}^T}{||\boldsymbol{b}||^2} = \frac{\boldsymbol{b}\boldsymbol{b}^T}{\boldsymbol{b}^T\boldsymbol{b}} = \frac{1}{9}\begin{bmatrix}1\\1\\2\end{bmatrix}\begin{bmatrix}1 & 1 & 2\end{bmatrix} = \frac{1}{9}\begin{bmatrix}1 & 2 & 2\\2 & 4 & 4\\2 & 4 & 4\end{bmatrix}$$

should we want to project $\boldsymbol{x} = [1, 1, 1]^T$ onto the one-dimensional subspace, that would look like

$$\pi_U(\boldsymbol{x}) = \boldsymbol{P}_\pi \boldsymbol{x} = \frac{1}{9}\begin{bmatrix}1 & 2 & 2\\2 & 4 & 4\\2 & 4 & 4\end{bmatrix}\begin{bmatrix}1\\1\\1\end{bmatrix} = \frac{1}{9}\begin{bmatrix}5\\10\\10\end{bmatrix} \in \text{span}[\begin{bmatrix}1\\2\\2\end{bmatrix}]$$

Now notice that application of $\boldsymbol{P}_\pi$ to $\pi_U(\boldsymbol{x})$ does not change anything:

$$\boldsymbol{P}_\pi^2\boldsymbol{x} = \frac{1}{9}\begin{bmatrix}1 & 2 & 2\\2 & 4 & 4\\2 & 4 & 4\end{bmatrix}\frac{1}{9}\begin{bmatrix}5\\10\\10\end{bmatrix} = \frac{1}{81}\begin{bmatrix}45\\90\\90\end{bmatrix} = \frac{1}{9}\begin{bmatrix}5\\10\\10\end{bmatrix}$$

This is expected since projections are defined to satisfy $\boldsymbol{P}_\pi^2\boldsymbol{x} = \boldsymbol{P}_\pi\boldsymbol{x}$ for all $\boldsymbol{x}$

Intuition for orthogonality implying minimum distance
Consider computing the distance:

$$d(v, U) := \min_{u \in U} ||v - u||$$

using the Pythagoras theorem, one has

$$||v - u||^2 = ||v - \pi(v)||^2 + ||u - \pi(v)||^2 \geq ||v - \pi(v)||^2$$

A.2.11 Orthogonal Projections II

Projection onto General Subspaces
Now we consider orthogonal projections of vectors $\boldsymbol{x} \in \mathbb{R}^n$ onto lower dimensional subspaces $U \subseteq \mathbb{R}^n$ with $\dim(U) = m \geq 1$:

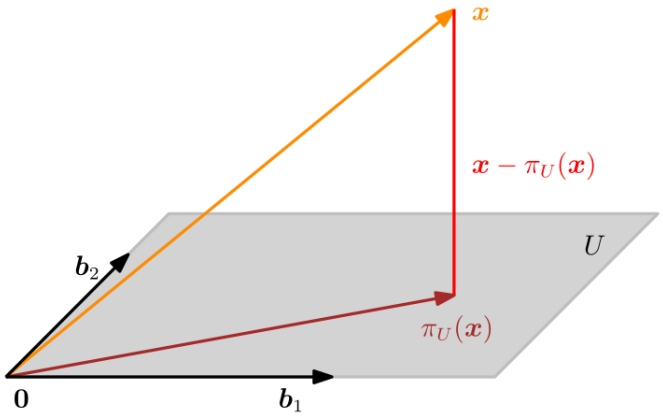


Figure: Projection onto a two-dimensional subspace U with basis $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$

Assume that $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m)$ is an ordered basis of U. Any projection $\pi_U(\boldsymbol{x})$ onto U is an element of U; therefore they can be represented as linear combinations of the basis vectors $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m$ of U, such that $\pi_U(\boldsymbol{x}) = \sum_{i=1}^{m} \lambda_i \boldsymbol{b}_i$.

Deriving the projection
Procedurally similar to the one-dimensional case:

1. **Determining coordinates of projection**
   We want the coordinates $\lambda_1, \ldots, \lambda_m$ of the projection (with respect to the basis of U), such that the linear combination

$$\pi_U(\boldsymbol{x}) = \sum_{i=1}^{m} \lambda_i \boldsymbol{b}_i = \boldsymbol{B}\boldsymbol{\lambda}$$
$$\text{where} \quad \boldsymbol{B} = [\boldsymbol{b_1}, \ldots, \boldsymbol{b_m}] \in \mathbb{R}^{n \times m}, \quad \boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_m]^T \in \mathbb{R}^m$$

   is closest to $\boldsymbol{x} \in \mathbb{R}^n$. Similar to the 1D case, this implies that the vector connecting $\pi_U(\boldsymbol{x}) \in U$ and $\boldsymbol{x} \in \mathbb{R}^n$ must be orthogonal to all basis vectors in U.
   (next page)

The vector connecting $\pi_U(\boldsymbol{x}) \in U$ and $\boldsymbol{x} \in \mathbb{R}^n$ must be orthogonal to all basis vectors in U—so we have m simultaneous conditions. Assuming the dot product as the inner product:

$$\begin{aligned}\langle \boldsymbol{b}_1, \boldsymbol{x} - \pi_U(\boldsymbol{x}) \rangle &= \boldsymbol{b}_1^T(\boldsymbol{x} - \pi_U(\boldsymbol{x})) = 0 \\ &\vdots \\ \langle \boldsymbol{b}_m, \boldsymbol{x} - \pi_U(\boldsymbol{x}) \rangle &= \boldsymbol{b}_m^T(\boldsymbol{x} - \pi_U(\boldsymbol{x})) = 0\end{aligned}$$

since $\pi_U(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{\lambda}$, we have

$$\begin{aligned}\boldsymbol{b}_1^T(\boldsymbol{x} - \boldsymbol{B}\boldsymbol{\lambda}) &= 0 \\ &\vdots \\ \boldsymbol{b}_m^T(\boldsymbol{x} - \boldsymbol{B}\boldsymbol{\lambda}) &= 0\end{aligned}$$

This can be written compactly as

$$\underbrace{\begin{bmatrix}\boldsymbol{b}_1^T \\ \vdots \\ \boldsymbol{b}_m^T\end{bmatrix}}_{m \times n} \underbrace{\left[\boldsymbol{x} - \boldsymbol{B}\boldsymbol{\lambda}\right]}_{n \times 1} = \boldsymbol{0} \iff \boldsymbol{B}^T(\boldsymbol{x} - \boldsymbol{B}\boldsymbol{\lambda}) = \boldsymbol{0}$$

$$\iff \boldsymbol{B}^T\boldsymbol{B}\boldsymbol{\lambda} = \boldsymbol{B}^T\boldsymbol{x}$$

This expression is called the *normal equation*. Since $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m$ are a basis of U and therefore linearly independent (since $\text{rk}(\boldsymbol{B})$ =$\text{rk}(\boldsymbol{B}^T)$ meaning B^TB can be seen as linear combinations of linearly independent vectors with linearly independent coefficients, making the result linearly independent (A.1.4)), $B^TB \in \mathbb{R}^{m \times m}$ is full rank and invertible. So we have

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

The matrix $\boldsymbol{B}^T\boldsymbol{B}$ will always be symmetric and positive semidefinite; symmetry is obvious, positive definiteness comes from $\boldsymbol{x}^T(\boldsymbol{B}^T\boldsymbol{B})\boldsymbol{x} = (\boldsymbol{B}\boldsymbol{x})^T(\boldsymbol{B}\boldsymbol{x}) \geq \boldsymbol{0}$. Invertibility of $\boldsymbol{B}^T\boldsymbol{B}$ comes from $\boldsymbol{B}$ being a basis and therefore full rank.

The matrix $(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T$ is also called the *psuedo-inverse* of $\boldsymbol{B}$.
(next page)

2. **Finding the projection point**
   We want the projection point $\pi_U(\boldsymbol{x})$; since $\pi_U(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{\lambda}$, we have

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

3. **Finding the projection matrix** We can immediately see that the projection matrix that solves $\boldsymbol{P}_\pi\boldsymbol{x} = \pi_U(\boldsymbol{x})$ would be

$$\boldsymbol{P}_\pi = \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T$$

   Notice that this general expression includes the 1D case as a special case, where if $\dim(U) = 1$, then $\boldsymbol{B}^T\boldsymbol{B} \in \mathbb{R}$ is a scalar so $\boldsymbol{P}_\pi = \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T = \frac{\boldsymbol{B}\boldsymbol{B}^T}{\boldsymbol{B}^T\boldsymbol{B}}$ (which is the projection matrix in the 1D case).

Example: 2D case
Considering a subspace $U = \text{span}[\begin{bmatrix}1\\1\\1\end{bmatrix}, \begin{bmatrix}0\\1\\2\end{bmatrix}] \subseteq \mathbb{R}^3$ and $\boldsymbol{x} = \begin{bmatrix}6\\0\\0\end{bmatrix} \in \mathbb{R}^3$, we want the coordinates $\boldsymbol{\lambda}$ of the projection of $\boldsymbol{x}$ onto U, the projection point $\pi_U(\boldsymbol{x})$, and the projection matrix $\boldsymbol{P}_\pi$.

First we see that the generating set of U is a basis (linear independence) and write the basis of U as a matrix $\boldsymbol{B} = \begin{bmatrix}1 & 0\\1 & 1\\1 & 2\end{bmatrix}$. Next we compute the matrix $\boldsymbol{B}^T\boldsymbol{B}$ and the vector $\boldsymbol{B}^T\boldsymbol{x}$:

$$\boldsymbol{B}^T\boldsymbol{B} = \begin{bmatrix}1 & 1 & 1\\0 & 1 & 2\end{bmatrix}\begin{bmatrix}1 & 0\\1 & 1\\1 & 2\end{bmatrix} = \begin{bmatrix}3 & 3\\3 & 5\end{bmatrix}, \quad \boldsymbol{B}^T\boldsymbol{x} = \begin{bmatrix}1 & 1 & 1\\0 & 1 & 2\end{bmatrix}\begin{bmatrix}6\\0\\0\end{bmatrix} = \begin{bmatrix}6\\0\end{bmatrix}$$

we solve the normal equation $\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{\lambda} = \boldsymbol{B}^T\boldsymbol{x}$ to find $\boldsymbol{\lambda}$:

$$\begin{bmatrix}3 & 3\\3 & 5\end{bmatrix}\begin{bmatrix}\lambda_1\\\lambda_2\end{bmatrix} = \begin{bmatrix}6\\0\end{bmatrix} \iff \boldsymbol{\lambda} = \begin{bmatrix}5\\-3\end{bmatrix}$$

The projection point $\pi_U(\boldsymbol{x})$ can be directly computed since $\pi_U(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{\lambda}$:

$$\pi_U(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{\lambda} = \begin{bmatrix}5\\2\\-1\end{bmatrix}$$

The projection matrix (for any $\boldsymbol{x} \in \mathbb{R}^3$) is given by

$$\boldsymbol{P}_\pi = \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T = \frac{1}{6}\begin{bmatrix}5 & 2 & -1\\2 & 2 & 2\\-1 & 2 & 5\end{bmatrix}$$

(next page)

Remarks
The projections $\pi_U(\boldsymbol{x})$ are still vectors in $\mathbb{R}^n$—they just lie in an m-dimensional subspace $U \subseteq \mathbb{R}^n$. To represent a projected point we only need m coordinates $\lambda_1, \ldots, \lambda_m$ with respect to the basis vectors $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m$ of U.

In vector spaces with general inner products, we have to pay attention when computing angles and distances, which are defined by means of the inner product.

Utility of Projections in unsolvable linear systems
Projections are useful in situations where we have a linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ without a solution, meaning $\boldsymbol{b}$ does not lie in the span of $\boldsymbol{A}$ (the column space of $\boldsymbol{A}$). Since we cannot find an exact solution we can find an *approximate solution* by finding the vector in the column space of $\boldsymbol{A}$ that is closest to $\boldsymbol{b}$—computing the orthogonal projection of $\boldsymbol{b}$ onto the column space of $\boldsymbol{A}$. (this solution is called the *least-squares solution* when the dot product is used as the inner product)

Utility of an Orthonormal Basis in finding projections
Should the subspace U that we are projecting onto have an orthonormal basis, notice that since $\boldsymbol{B}^T\boldsymbol{B} = \boldsymbol{I}$ the projection equation simplifies greatly to

$$\pi_U(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{B}^T\boldsymbol{x}$$

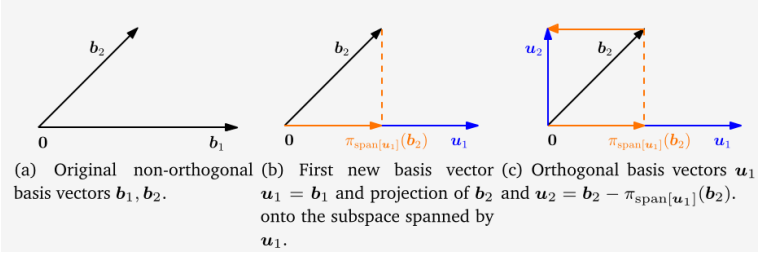
we no longer have to compute $(\boldsymbol{B}^T\boldsymbol{B})^{-1}$, which saves computation time.

A.2.12 Gram-Schmidt Orthogonalisation

The Gram-Schmidt method uses projections to constructively transform any basis $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)$ of an n-dimensional vector space V into an orthogononal or orthonormal basis $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ of V. An orthogonal basis is iteratively constructed from any basis as follows:

$$\begin{aligned}\boldsymbol{u}_1 &:= \boldsymbol{b}_1 \\ \boldsymbol{u}_k &:= \boldsymbol{b}_k - \pi_{\text{span}[\boldsymbol{u}_1,\ldots,\boldsymbol{u}_{k-1}]}(\boldsymbol{b}_k), \quad k = 2, \ldots, n\end{aligned}$$

The kth basis vector $\boldsymbol{b}_k$ is projected onto the subspace spanned by the first $k-1$ constructed orthogonal vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{k-1}$. This projection is then subtracted from $\boldsymbol{b}_k$ and yields a vector $\boldsymbol{u}_k$ that is orthogonal to the $(k-1)$-dimensional subspace spanned by $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{k-1}$.



(a) Original non-orthogonal basis vectors $\boldsymbol{b}_1, \boldsymbol{b}_2$.

(b) First new basis vector $\boldsymbol{u}_1 = \boldsymbol{b}_1$ and projection of $\boldsymbol{b}_2$ onto the subspace spanned by $\boldsymbol{u}_1$.

(c) Orthogonal basis vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2 = \boldsymbol{b}_2 - \pi_{\text{span}[\boldsymbol{u}_1]}(\boldsymbol{b}_2)$.

Repeating this procedure for all n original basis vectors $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)$ yields an orthogonal basis $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ of V. If we normalise the $\boldsymbol{u}_k$ we then obtain an orthonormal basis where $||\boldsymbol{u}_k|| = 1$ for all $k = 1, \ldots, n$.

Example
Consider a basis $(\boldsymbol{b}_1, \boldsymbol{b}_2)$ of $\mathbb{R}^2$, where

$$\boldsymbol{b}_1 = \begin{bmatrix}2\\0\end{bmatrix}, \quad \boldsymbol{b}_2 = \begin{bmatrix}1\\1\end{bmatrix}$$

Using the Gram-Schmidt method we construct an orthogonal basis $(\boldsymbol{u}_1, \boldsymbol{u}_2)$ of $\mathbb{R}^2$ as follows (assuming the dot product as the inner product):

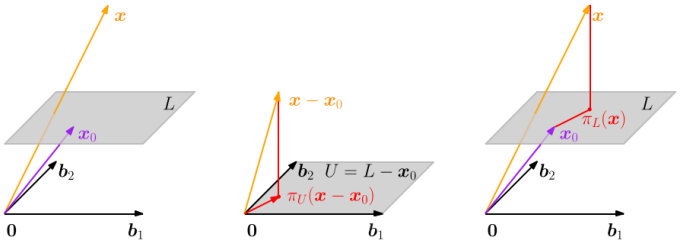
$$\boldsymbol{u}_1 = \boldsymbol{b}_1$$

$$\boldsymbol{u}_2 = \boldsymbol{b}_2 - \pi_{\text{span}[\boldsymbol{u}_1]}(\boldsymbol{b}_2) = \boldsymbol{b}_2 - \frac{\boldsymbol{u}_1\boldsymbol{u}_1^T}{||\boldsymbol{u}_1||^2}\boldsymbol{b}_2 = \begin{bmatrix}1\\1\end{bmatrix} - \begin{bmatrix}1 & 0\\0 & 0\end{bmatrix}\begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}0\\1\end{bmatrix}$$

see that $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are orthogonal: $\boldsymbol{u}_1^T\boldsymbol{u}_2 = 0$

A.2.13 Projection onto Affine Subspaces

Given an affine space $L = \boldsymbol{x}_0 + U$, where $\boldsymbol{b}_1, \boldsymbol{b}_2$ are basis vectors of U, we want to determine the orthogonal projection $\pi_L(\boldsymbol{x})$ of some $\boldsymbol{x}$ onto L.



(a) Setting.

(b) Reduce problem to projection π_U onto vector subspace.

(c) Add support point back in to get affine projection π_L.

An approach would be to subtract the support point $\boldsymbol{x}_0$ from $\boldsymbol{x}$ and from L, since $L - \boldsymbol{x}_0 = U$. We then take the projection $\pi_U(\boldsymbol{x} - \boldsymbol{x}_0)$, then translate the projection back to L by adding $\boldsymbol{x}_0$:

$$\pi_L(\boldsymbol{x}) = \boldsymbol{x}_0 + \pi_U(\boldsymbol{x} - \boldsymbol{x}_0)$$

Notice that the distance of $\boldsymbol{x}$ from the affine space L is equal to the distance of $\boldsymbol{x} - \boldsymbol{x}_0$ from U:

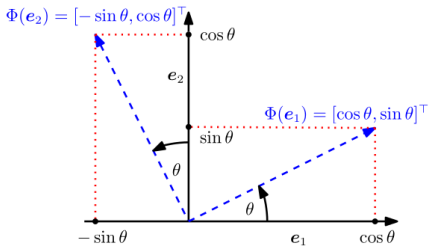
$$\begin{aligned} d(\boldsymbol{x}, L) &= ||\boldsymbol{x} - \pi_L(\boldsymbol{x})|| = ||\boldsymbol{x} - (\boldsymbol{x}_0 + \pi_U(\boldsymbol{x} - \boldsymbol{x}_0))|| \\ &= d(\boldsymbol{x} - \boldsymbol{x}_0, \pi_U(\boldsymbol{x} - \boldsymbol{x}_0)) = d(\boldsymbol{x} - \boldsymbol{x}_0, U) \end{aligned}$$

A.2.14 Rotations

Orthogonal transformation matrices (A.2.7) preserve length and angle; some of these matrices describe rotations.

Rotations in $\mathbb{R}^2$

A *rotation* is a linear mapping (an automorphism) that rotates a plane by an angle θ about the origin. Considering the standard basis $\left\{\boldsymbol{e}_1 = \begin{bmatrix}1\\0\end{bmatrix}, \boldsymbol{e}_2 = \begin{bmatrix}0\\1\end{bmatrix}\right\}$, say we want to rotate this coordinate system by an angle θ:



Rotations Φ are linear mappings so we can express them by a *rotation matrix* $\boldsymbol{R}(\theta)$. We determine the coordinates of the rotated axes (the image of Φ) with respect to the standard basis:

$$\Phi(\boldsymbol{e}_1) = \begin{bmatrix}\cos\theta\\ \sin\theta\end{bmatrix}, \quad \Phi(\boldsymbol{e}_2) = \begin{bmatrix}-\sin\theta\\ \cos\theta\end{bmatrix}$$

Therefore the rotation that performs the basis change into the rotated coordinates $\boldsymbol{R}(\theta)$ is given as

$$\boldsymbol{R}(\theta) = \begin{bmatrix}\Phi(\boldsymbol{e}_1) & \Phi(\boldsymbol{e}_2)\end{bmatrix} = \begin{bmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{bmatrix}$$

(next page)

Rotations in $\mathbb{R}^3$

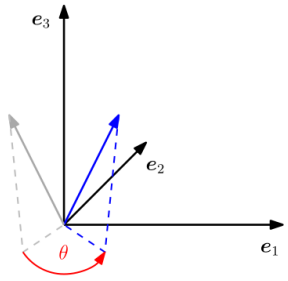
In contrast to the $\mathbb{R}^2$ case, in $\mathbb{R}^3$ we can rotate any two-dimensional plane about a one-dimensional axis. The easiest way to specify the general rotation matrix is to specify the images of the standard basis, and making sure they are orthonormal to each other (so we get an orthogonal matrix which preserves length and angle).

We define a 'counterclockwise' rotation about an axis as a rotation about that axis when we look at it 'head on, from the end toward the origin'. Therefore in $\mathbb{R}^3$ there are three (planar) rotations about the three standard basis vectors:

- Rotation about the $\boldsymbol{e}_3$-axis:

$$\boldsymbol{R}_3(\theta) = \begin{bmatrix} \Phi(\boldsymbol{e}_1) & \Phi(\boldsymbol{e}_2) & \Phi(\boldsymbol{e}_3) \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

  The $\boldsymbol{e}_3$ coordinate is fixed (its image is therefore the same), and counterclockwise rotation is performed in the $\boldsymbol{e}_1\boldsymbol{e}_2$ plane. We look at $\boldsymbol{e}_3$ from its 'tip' toward the origin:



- Rotation about the $\boldsymbol{e}_1$-axis (counterclockwise):

$$\boldsymbol{R}_1(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}$$

- Rotation about the $\boldsymbol{e}_2$-axis (clockwise):

$$\boldsymbol{R}_2(\theta) = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix}$$

(next page)

Rotations in n Dimensions
The generalisation of rotations from 2D to 3D to n-dimensional Euclidean vector spaces can be intuitively described as fixing n-2 dimensions and restricting the rotation to a two-dimensional plane in the n-dimensional space; here we describe the Givens Rotation:

Givens rotation
Let V be an n-dimensional Euclidean vector space and $\Phi : V \to V$ an automorphism with transformation matrix

$$\boldsymbol{R}_{ij}(\theta) := \begin{bmatrix} \boldsymbol{I}_{i-1} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \cos\theta & \mathbf{0} & -\sin\theta & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{I}_{j-i-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sin\theta & \mathbf{0} & \cos\theta & \mathbf{0} \\ \mathbf{0} & \cdots & \cdots & \mathbf{0} & \boldsymbol{I}_{n-j} \end{bmatrix}$$

for $1 \leq i < j \leq n$ and $\theta \in \mathbb{R}$. Then $\boldsymbol{R}_{ij}$ is called a *Givens rotation.* Essentially, $\boldsymbol{R}_{ij}$ is the identity matrix $\boldsymbol{I}_n$ with

$$r_{ii} = \cos\theta, \quad r_{ij} = -\sin\theta, \quad r_{ji} = \sin\theta, \quad r_{jj} = \cos\theta$$

Properties of Rotations Most properties of rotations come from them being orthogonal matrices:

- Rotations preserve distances, meaning $||\boldsymbol{x} - \boldsymbol{y}|| = ||\boldsymbol{R}_\theta \boldsymbol{x} - \boldsymbol{R}_\theta \boldsymbol{y}||$.
- Rotations preserve angles, meaning the angle between $\boldsymbol{R}_\theta \boldsymbol{x}$ and $\boldsymbol{R}_\theta \boldsymbol{y}$ equals the angle between $\boldsymbol{x}$ and $\boldsymbol{y}$.
- Rotations in three or more dimensions are generally non commutative. Only in two dimensions are vector rotations commutative, this means $\boldsymbol{R}(\phi)\boldsymbol{R}(\theta) = \boldsymbol{R}(\phi)\boldsymbol{R}(\theta)$ for all $\phi, \theta \in [0, 2\pi]$. They form an Abelian group (with multiplication) only if they rotate about the same point.

A.3 Matrix Decompositions

A.3.1 Properties of the determinant

Geometrically the determinant of a matrix represents the signed volume of the parallelepiped formed by the columns of that matrix.

For $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ the determinant exhibits the following properties:

- The determinant of a matrix product is the product of the corresponding determinants, meaning $\det(\boldsymbol{AB}) = \det(\boldsymbol{A})\det(\boldsymbol{B})$.
- Determinants are invariant to transposition, meaning $\det(\boldsymbol{A}) = \det(\boldsymbol{A}^T)$.
- If $\boldsymbol{A}$ is invertible/regular, then $\det(\boldsymbol{A}^{-1}) = \frac{1}{\det(\boldsymbol{A})}$.
- Similar matrices (A.1.11) possess the same determinant. Therefore for a linear mapping $\Phi : V \to V$ all transformation matrices $\boldsymbol{A}_\Phi$ of Φ have the same determinant; the determinant is invariant to the choice of basis for a linear mapping.
- Adding a multiple of a column/row to another one does not change $\det(\boldsymbol{A})$.
- Multiplication of a column/row with $\lambda \in \mathbb{R}$ scales $\det(\boldsymbol{A})$ by λ. In particular, $\det \lambda\boldsymbol{A} = \lambda^n \det(\boldsymbol{A})$.
- Swapping two columns changes the sign of $\det(\boldsymbol{A})$.

Notice that because of the last three properties, we can use Gaussian elimination to compute $\det(\boldsymbol{A})$ by bringing $\boldsymbol{A}$ to row-echelon form. We can stop Gaussian elimination when we have $\boldsymbol{A}$ in upper triangular form—so that the determinant is just the product of the diagonal elements.

Determinant and Rank
Theorem: *A square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times b}$ has det($\boldsymbol{A}$) $\neq 0$ if and only if rk($\boldsymbol{A}$) $= n$. In other words, $\boldsymbol{A}$ is invertible if and only if it is full rank.*

A.3.2 Trace

The *trace* of a square matrix $\boldsymbol{A}^{n\times n}$ is defined as

$$\mathrm{tr}(A) := \sum_{i=1}^{n} a_{ii}$$

essentially the sum of the elements on the main diagonal.

The trace satisfies the following properties:

- $\mathrm{tr}(\boldsymbol{A}+\boldsymbol{B}) = \mathrm{tr}(A) + \mathrm{tr}(\boldsymbol{B})$ for $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$
- $\mathrm{tr}(\alpha\boldsymbol{A}) = \alpha\mathrm{tr}(\boldsymbol{A})$, $\alpha \in \mathbb{R}$ for $\boldsymbol{A} \in \mathbb{R}^{n\times n}$.
- $\mathrm{tr}(\boldsymbol{I}_n) = n$
- $\mathrm{tr}(\boldsymbol{AB}) = \mathrm{tr}(\boldsymbol{BA})$ for $\boldsymbol{A} \in \mathbb{R}^{n\times k}, \boldsymbol{B}^{k\times n}$

Invariance under cyclic permutation
Notice that the last property can be extended— the trace is invariant under cyclic permutation of matrix products:

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

For matrices $\boldsymbol{A} \in \mathbb{R}^{n\times k}, \boldsymbol{K} \in \mathbb{R}^{k\times l}, \boldsymbol{L} \in \mathbb{R}^{l\times n}$. This property generalises to products of an arbritrary number of matrices. A special case of this should be noted: for two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

$$\mathrm{tr}(\boldsymbol{xy}^T) = \mathrm{tr}(\boldsymbol{y}^T\boldsymbol{x}) = \boldsymbol{y}^T\boldsymbol{x} \in \mathbb{R}$$

(next page)

Trace, Linear mappings, and Basis changes
Given a linear mapping $\Phi : V \to V$, where V is a vector space, we define the trace of this map by using the trace of the matrix representation of Φ; for a given basis of V, we can describe Φ by means of the transformation matrix $\boldsymbol{A}$, where the trace of Φ is the trace of $\boldsymbol{A}$.

For a different basis of V, it holds that the corresponding transformation matrix $\boldsymbol{B}$ of Φ can be obtained by a basis change of the form $\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S}$ for suitable $\boldsymbol{S}$ (A.1.10). For the corresponding trace of Φ, this means

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

While matrix representations of linear mappings are basis dependent, the trace of a linear mapping Φ is independent of the basis.

A.3.3 Eigenvalues, Eigenvectors, Characteristic Polynomial

Definition of Characteristic polynomial
For $\lambda \in \mathbb{R}$ and a square matrix $A \in \mathbb{R}^{n\times n}$

$$\begin{aligned} p_{\boldsymbol{A}}(\lambda) &:= \det(\boldsymbol{A} - \lambda \boldsymbol{I}) \\ &= c_0 + c_1\lambda + c_2\lambda^2 + \ldots + c_{n-1}\lambda^{n-1} + (-1)^n\lambda^n \end{aligned}$$

$c_0, \ldots, c_{n-1} \in \mathbb{R}$ is the characteristic polynomial of $\boldsymbol{A}$. In particular,

$$\begin{aligned} c_0 &= \det(\boldsymbol{A}) \\ c_{n-1} &= (-1)^{n-1}\mathrm{tr}(\boldsymbol{A}) \end{aligned}$$

Motivation for the characteristic polynomial comes from its utility in computing eigenvalues and eigenvectors.

Eigenvectors and Eigenvalues
Definition: Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be a square matrix. Then $\lambda \in \mathbb{R}$ is an *eigenvalue* of $\boldsymbol{A}$ and $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$ is the corresponding *eigenvector* of $\boldsymbol{A}$ if

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

the above is called the *eigenvalue equation*. The following statements are equivalent:

- λ is an eigenvalue of $\boldsymbol{A} \in \mathbb{R}^{n\times n}$.
- There exists an $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$ with $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{\lambda}\boldsymbol{x}$; or equivalently, $(\boldsymbol{A} - \lambda \boldsymbol{I}_n)\boldsymbol{x} = \boldsymbol{0}$ can be solved non trivially. (If an eigenvalue exists then an eigenvector exists.)
- $\mathrm{rk}(\boldsymbol{A} - \lambda \boldsymbol{I}_n) < n$ (since there are nontrivial solutions).
- $\det(\boldsymbol{A} - \lambda \boldsymbol{I}_n) = 0$

Notice this motivates the necessity for the characteristic equation— $\lambda \in \mathbb{R}$ is an eigenvalue of $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ if and only if λ is a root of the characteristic polynomial $p_{\boldsymbol{A}}(\lambda)$ of $\boldsymbol{A}$.

Collinearity, Codirection—definition and significance
Two vectors that point in the same direction are called *codirected*. Two vectors are *collinear* if they point in the same or opposite direction.

If $\boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}$ associated with eigenvalue λ, then for any $c \in \mathbb{R}\backslash\{\boldsymbol{0}\}$ it holds that $c\boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}$ with the same eigenvalue since

$$\boldsymbol{A}(c\boldsymbol{x}) = c\boldsymbol{A}(\boldsymbol{x}) = c\lambda(\boldsymbol{x}) = \lambda(c\boldsymbol{x})$$

All vectors that are collinear to $\boldsymbol{x}$ are also eigenvectors of $\boldsymbol{A}$.
(next page)

Algebraic multiplicity
Let a square matrix $\boldsymbol{A}$ have an eigenvalue λ. The *algebraic multiplicity* of λ is the number of times the root appears in the characteristic polynomial.

Eigenspace and Eigenspectrum
For $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ the set of all eigenvectors of $\boldsymbol{A}$ associated with eigenvalue λ spans a subspace of $\mathbb{R}^n$ called the *eigenspace* of $\boldsymbol{A}$ with respect to λ and is denoted by E_λ. The set of eigenvalues of $\boldsymbol{A}$ is called the *eigenspectrum/spectrum* of $\boldsymbol{A}$.

If λ is an eigenvalue of $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, then the corresponding eigenspace is the solution space of $(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{x} = \boldsymbol{0}$.

Notable properties

- A matrix $\boldsymbol{A}$ and its transpose $\boldsymbol{A}^T$ possess the same eigenvalues, but not necessarily the same eigenvectors (for intuition, recall that the determinant is invariant with respect to transposition)
- The eigenspace E_λ is the nullspace of $\boldsymbol{A} - \lambda\boldsymbol{I}$ since

$$\begin{aligned}\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x} &\iff \boldsymbol{A}\boldsymbol{x} - \lambda\boldsymbol{x} = \boldsymbol{0} \\ &\iff (\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{x} = \boldsymbol{0} \iff \boldsymbol{x} \in \ker(\boldsymbol{A} - \lambda\boldsymbol{I})\end{aligned}$$

- Similar matrices possess the same eigenvalues. Therefore a linear mapping Φ has eigenvalues that are independent of the choice of basis of its transformation matrix. (this makes eigenvalues, together with the determinant and the trace, key characteristic parameters of a linear mapping as they are all invariant under basis change)
- Symmetric, positive definite matrices always have positive, real eigenvalues (A.3.4).

The case of the identity matrix
The identity matrix $\boldsymbol{I} \in \mathbb{R}^{n\times n}$ has characteristic polynomial $p_I(\lambda) = (det)(\boldsymbol{I} - \lambda\boldsymbol{I}) = (1-\lambda)^n = 0$, which only has one eigenvalue $\lambda = 1$ that occurs n times.

Moreover, $\boldsymbol{I}\boldsymbol{x} = \lambda\boldsymbol{x} = 1\boldsymbol{x}$ holds for all vector $\boldsymbol{x} \in \mathbb{R}\backslash\{\boldsymbol{0}\}$; because of this the sole eigenspace E_1 of the identity matrix spans n dimensions, and all n standard basis vectors of $\mathbb{R}^n$ are eigenvectors of $\boldsymbol{I}$.
(next page)

Example: Computing Eigenvectors, Eigenvalues, and Eigenspaces
Consider finding the eigenvalues and eigenvectors of the 2×2 matrix

$$\boldsymbol{A} = \begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix}$$

Step 1: Characteristic polynomial
The necessity for the characteristic polynomial comes from the idea that there will be a vector such that $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$, meaning $(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{x} = \boldsymbol{0}$. For $\boldsymbol{x} \neq \boldsymbol{0}$ this requires that the kernel of $\boldsymbol{A} - \lambda\boldsymbol{I}$ contains more than just $\boldsymbol{0}$; this means that $\boldsymbol{A} - \lambda\boldsymbol{I}$ is not invertible and therefore that $\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0$. Hence we need to compute the roots of the characteristic 0polynomial to find the eigenvalues. This is given by

$$\begin{aligned} p_{\boldsymbol{A}}(\lambda) &= \det(\boldsymbol{A} - \lambda\boldsymbol{I}) \\ &= \det\left(\begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right) = \begin{vmatrix} 4-\lambda & 2 \\ 1 & 3-\lambda \end{vmatrix} \\ &= (4-\lambda)(3-\lambda) - 2 \cdot 1 \end{aligned}$$

Factorising, we obtain

$$p_{\boldsymbol{A}}(\lambda) = (2-\lambda)(5-\lambda)$$

with that we have the eigenvalues $\lambda_1 = 2$ and $\lambda_2 = 5$.

Step 2: Eigenvectors and Eigenspaces
We find the eigenvectors by solving for $\boldsymbol{x}$ in the aforementioned system:

$$\begin{bmatrix} 4-\lambda & 2 \\ 1 & 3-\lambda \end{bmatrix} \boldsymbol{x} = \boldsymbol{0}$$

For $\lambda = 5$ we obtain

$$\begin{bmatrix} 4-5 & 2 \\ 1 & 3-5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1 & 2 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \boldsymbol{0}$$

Solving this homogeneous system we obtain a solution space

$$E_5 = \text{span}[\begin{bmatrix} 2 \\ 1 \end{bmatrix}]$$

This eigenspace is one dimensional since it contains a single basis vector. We analagously solve for the eigenvector of $\lambda = 2$ to get the eigenspace

$$E_2 = \text{span}[\begin{bmatrix} 1 \\ -1 \end{bmatrix}]$$

The two eigenspaces E_5 and E_2 in this case are one-dimensional as they are each spanned by a single vector. However in other cases we may have multiple identical eigenvalues and the eigenspace may have more than one dimension.

A.3.4 Symmetric matrices always have real eigenvalues

Let $(\lambda, \boldsymbol{v})$ be any eigenpair of $\boldsymbol{A}$. For a symmetric matrix, $\boldsymbol{A} = \boldsymbol{A}^T$; so (for the dot product)

$$\langle \boldsymbol{A}\boldsymbol{v}, \boldsymbol{A}\boldsymbol{v} \rangle = (\boldsymbol{A}\boldsymbol{v})^T \boldsymbol{A}\boldsymbol{v} = (\lambda \boldsymbol{v})^T \lambda \boldsymbol{v} = \lambda^2 ||\boldsymbol{v}||^2$$

Since

$$\lambda^2 = \frac{\langle \boldsymbol{A}\boldsymbol{v}, \boldsymbol{A}\boldsymbol{v} \rangle}{||\boldsymbol{v}||^2}$$

is a real nonnegative number, λ must be real.

A.3.5 Eigenvalues and Eigenvectors II

Geometric Multiplicity

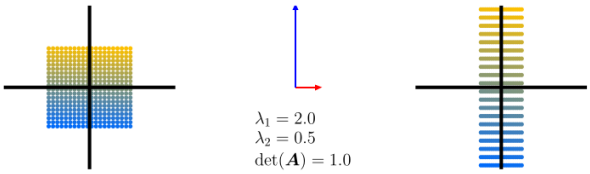
Let λ_i be an eigenvalue of a square matrix $\boldsymbol{A}$. Then the *geometric multiplicity* of λ_i is the number of linearly independent eigenvectors associated with λ_i—the dimensionality of the eigenspace spanned by the eigenvectors associated with λ_i.

A specific eigenvalue's geometric multiplicity must be at lease one because every eigenvalue has at least one associated eigenvector. An eigenvalue's geometric multiplicity cannot exceed its algebraic multiplicity, but it may be lower.

Graphical intuition in two dimensions

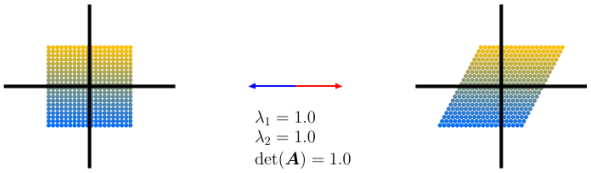
Geometrically, the eigenvector corresponding to a nonzero eigenvalue points in a direction that is stretched by the linear mapping. The eigenvalue is a factor by which it is stretched. For intuition we present a few examples:

- $\boldsymbol{A}_1 = \begin{bmatrix} 1/2 & 0 \\ 0 & 2 \end{bmatrix}$:



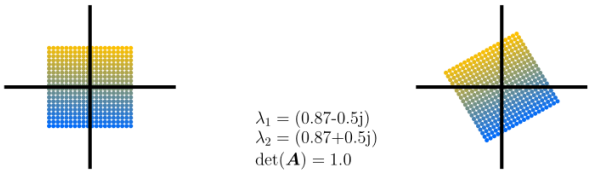
The direction of the two eigenvectors correspond to the canonical basis vectors in $\mathbb{R}^2$. The vertical axis is extended by a factor of 2 (eigenvalue $\lambda_1 = 2$), and the horizontal axis is compressed by a factor of $1/2$ (eigenvalue $\lambda_2 = 1/2$). The result is area preserving (since $\det(\boldsymbol{A}_1) = 1 = 2 \cdot 1/2$).
(next page)

- $\boldsymbol{A}_2 = \begin{bmatrix} 1 & 1/2 \\ 0 & 1 \end{bmatrix}$ corresponds to a shear mapping (it shears the points along the horizontal axis to the right if they are on the positive half of the vertical axis and to the left vice versa):



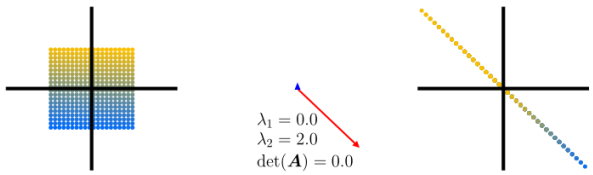
This mapping is also area preserving (notice the determinant). The eigenvalue $\lambda_1 = 1 = \lambda_2$ is repeated and the eigenvectors are collinear. This indicates that the mapping only acts along one direction (in this case the horizontal axis).

- $\boldsymbol{A}_3 = \begin{bmatrix} \cos(\pi/6) & -\sin(\pi/6) \\ \sin(\pi/6) & \cos(\pi/6) \end{bmatrix} = \frac{1}{2}\begin{bmatrix} \sqrt{3} & -1 \\ 1 & \sqrt{3} \end{bmatrix}$



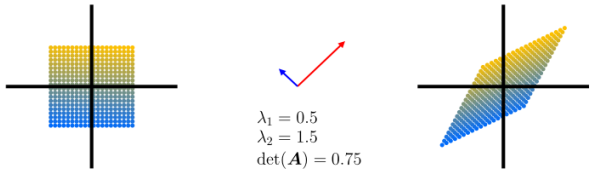
This matrix rotates the points by $\frac{\pi}{6}$rad$= 30°$ counter-clockwise and has only complex eigenvalues; this reflects the fact that the matrix represents a rotation mapping (and therefore doesn't have eigenvectors). Since a rotation is volume preserving, the determinant is 1.
(next page)

- $\boldsymbol{A_1} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$



represents a mapping that collapses a two-dimensional domain onto one domain. Since one eigenvalue is 0, the space in the direction of the corresponding eigenvector collapses, while the orthgonal eigenvector stretches space by a factor $\lambda = 2$; therefore the area of the image is 0. (notice the determinant is 0)

- $\boldsymbol{A}_5 = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}$



is a stretch-and-shear mapping that scales space by 75% (since $|\det(\boldsymbol{A}_5)| = \frac{3}{4}$) it stretches space along the eigenvector of $\lambda = 2$ by factor 1.5 and compresses it along the orthogonal eigenvector by factor 0.5.

A.3.6 Eigenvalues and Eigenvectors III

Linear independence of eigenvectors
Theorem: *The eigenvectors $x_1, \ldots, x_n$ of a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ with n distinct eigenvalues $\lambda_1, \ldots, \lambda_n$ are linearly independent.*

In essence, the eigenvectors of a matrix with n distinct eigenvalues form a basis of $\mathbb{R}^n$.

Defective matrices
A square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is *defective* if it possesses fewer than n linearly independent eigenvectors.

A non-defective matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ does not necessarily require n distinct eigenvalues, but rather that the eigenvectors form a basis of $\mathbb{R}^n$. Looking at the eigenvectors of a defective matrix, it follows that the sum of the dimensions of the eigenspaces is less than n.

Specifically, a defective matrix has at least one eigenvalue λ_i, with an algebraic multiplicity $m > 1$ but a geometric multiplicity less than m. This must be the case since a defective matrix cannot have n distinct eigenvalues, as distinct eigenvalues have distinct eigenvectors (as per the theorem regarding linear independence of eigenvectors).

A.3.7 Symmetry and Positive definiteness of $\boldsymbol{A}^T\boldsymbol{A}$, Spectral Theorem

Symmetry and Positive definiteness of $\boldsymbol{A}^T\boldsymbol{A}$
Theorem: *Given a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, we can always obtain a symmetric, positive semidefinite matrix $\boldsymbol{S} \in \mathbb{R}^{n\times n}$ by defining*

$$\boldsymbol{S} := \boldsymbol{A}^T\boldsymbol{A}$$

See that symmetry comes from

$$\boldsymbol{S} = \boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{A}^T(\boldsymbol{A}^T)^T = (\boldsymbol{A}^T\boldsymbol{A})^T = \boldsymbol{S}^T$$

(recall $(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$) (or notice that the element in position ij of the product will always be the same as the element in ji).

Positive definiteness comes from

$$\boldsymbol{x}^T\boldsymbol{S}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = (\boldsymbol{x}^T\boldsymbol{A}^T)(\boldsymbol{A}\boldsymbol{x}) = (\boldsymbol{A}\boldsymbol{x}^T)(\boldsymbol{A}\boldsymbol{x}) \geq 0$$

because the dot product computes a sum of squares which are non-negative.
(next page)

Spectral Theorem
If $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is symmetric, there exists an orthonormal basis of the corresponding vector space V consisting of eigenvectors of $\boldsymbol{A}$, and each eigenvalue is real.

Example
Consider the matrix

$$\boldsymbol{A} = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix}$$

The characteristic polynomial of $\boldsymbol{A}$ is

$$p_{\boldsymbol{A}}(\lambda) = -(\lambda - 1)^2(\lambda - 7)$$

so we obtain eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 7$, where λ_1 is a repeat eigenvalue. Following the standard procedure for obtaining eigenvectors, we get the eigenspaces

$$E_1 = \text{span}[\underbrace{\begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}}_{=:\boldsymbol{x}_1}, \underbrace{\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}}_{=:\boldsymbol{x}_2}], \quad E_7 = \text{span}[\underbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}}_{\boldsymbol{x}_3}]$$

Of the eigenvectors we found, $\boldsymbol{x}_3$ is orthogonal to both $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, but $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are not orthogonal to each other. The spectral theorem states that there exists an orthogonal basis—we can construct one.

To construct such a basis we can exploit the fact that $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are eigenvectors associated with the same eigenvalue λ; it holds for any $\alpha, \beta \in \mathbb{R}$ that

$$\boldsymbol{A}(\alpha\boldsymbol{x}_1 + \beta\boldsymbol{x}_2) = \boldsymbol{A}\boldsymbol{x}_1\alpha + \boldsymbol{A}\boldsymbol{x}_2\beta = \lambda(\alpha\boldsymbol{x}_1 + \beta\boldsymbol{x}_2)$$

Meaning that any linear combination of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is also an eigenvector of $\boldsymbol{A}$ associated with λ. We can use the Gram-Schmidt method, since it uses linear combinations (projections), to iteratively construct an orthogonal/orthonormal basis. Therefore even if $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are not orthogonal, applying the Gram-Schmidt algorithm:

$$\boldsymbol{x}_1' = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{x}_2' = \frac{1}{2}\begin{bmatrix} -1 \\ -1 \\ 2 \end{bmatrix}$$

which are orthogonal to each other and $\boldsymbol{x}_3$.

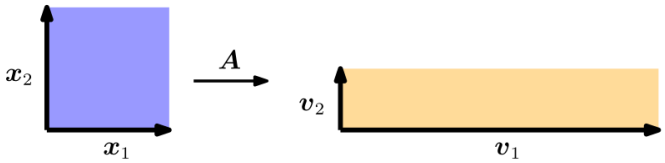
A.3.8 Determinant, Trace, and Eigenvalues

Eigenvalues and the Determinant
Theorem: *The determinant of a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is the product of its eigenvalues:*

$$\det(\boldsymbol{A}) = \prod_{i=1}^{n} \lambda_i$$

where $\lambda_i \in \mathbb{C}$ are (possibly repeated) eigenvalues of $\boldsymbol{A}$.

For geometric intuition, consider a matrix $\boldsymbol{A} \in \mathbb{R}^{2\times 2}$ that possesses two linearly independent eigenvectors $\boldsymbol{x}_1, \boldsymbol{x}_2$, assume that $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ are an ONB of $\mathbb{R}^2$ so that they are orthogonal and the area of the square they span is 1:



We know that the determinant computes the change of the unit square under the transformation $\boldsymbol{A}$. In this example, we can compute the change of the area explicitly; where mapping the eigenvectors using $\boldsymbol{A}$ gives us vectors $\boldsymbol{v}_1 = \boldsymbol{A}\boldsymbol{x}_1 = \lambda\boldsymbol{x}_1$ and $\boldsymbol{v}_2 = \boldsymbol{A}\boldsymbol{x}_2 = \lambda\boldsymbol{x}_2$—scaled versions of the initial eigenvectors, scaled by the eigenvalues. Now see that the the area of the rectangle that they now span is the absolute value of the product of the eigenvalues (because $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are orthonormal); the area is $|\lambda_1\lambda_2|$. (this is not a rigourous proof by any means, just some intuition)

Trace and Eigenvalues
Theorem: *The trace of a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is the sum of its eigenvalues*:

$$\mathrm{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} \lambda_i$$

where $\lambda_i \in \mathbb{C}$ are (possibly repeated eigenvalues of $\boldsymbol{A}$.

This can be proved by comparing coefficients with the equation for the characteristic polynomial.

A.3.9 Cholesky Decomposition

With positive real numbers, we have a square-root operation that gives us a decomposition into identical components. For symmetric, positive definite matrices, the *Cholesky decomposition/factorisation* is one of a number of operations that provides a square-root equivalent operation that is useful in practice.

Theorem: *A symmetric, positive definite matrix* $\boldsymbol{A}$ *can be factorised into a product* $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^T$*, where* $\boldsymbol{L}$ *is a lower triangular matrix with positive diagonal elements:*

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} l_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ l_{n1} & \cdots & l_{nn} \end{bmatrix} \begin{bmatrix} l_{11} & \cdots & l_{n1} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & l_{nn} \end{bmatrix}$$

$\boldsymbol{L}$ *is called the Cholesky factor of* $\boldsymbol{A}$ *and* $\boldsymbol{L}$ *is unique.*

3×3 case
Consider a symmetric, positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{3\times 3}$; we are interested in finding its Cholesky factorisation $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^T$:

$$\begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{21} & a_{22} & a_{32} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \boldsymbol{L}\boldsymbol{L}^T = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix}$$

Multplying out the right side yields

$$\begin{bmatrix} l_{11}^2 & l_{21}l_{11} & l_{31}l_{11} \\ l_{21}l_{11} & l_{21}^2 + l_{22}^2 & l_{31}l_{21} + l_{32}l_{22} \\ l_{31}l_{11} & l_{31}l_{21} + l_{32}l_{22} & l_{31}^2 + l_{32}^2 + l_{33}^2 \end{bmatrix}$$

Comparing the left and right hand sides we can see the diagonal elements follow a pattern:

$$\begin{aligned} a_{11} = l_{11}^2 &\iff l_{11} = \sqrt{a_{11}} \\ a_{22} = l_{21}^2 + l_{22}^2 &\iff l_{22} = \sqrt{a_{22} - l_{21}^2} \\ a_{33} = l_{31}^2 + l_{32}^2 + l_{33}^2 &\iff l_{33} = \sqrt{a_{33} - (l_{31}^2 + l_{32}^2)} \end{aligned}$$

Equations for the elements below the diagonal can also be teased out: (apparently with a repeating pattern)

$$l_{21} = \frac{1}{l_{11}} a_{21}, \quad l_{31} = \frac{1}{l_{11}} a_{31}, \quad l_{32} = \frac{1}{l_{22}}(a_{32} - l_{31}l_{21})$$

With that we have the Cholesky decomposition (for any symmetric, positive semidefinite 3×3 matrix); the key point here is that given $\boldsymbol{A}$, we can backward calculate all the components of $\boldsymbol{L}$.

A.3.10 Eigendecomposition and Diagonalisation

Definition and Significance of Diagonal matrices
A *diagonal matrix* is a matrix that has value zero on all the off-diagonal elements:

$$\boldsymbol{D} = \begin{bmatrix} c_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & c_n \end{bmatrix}$$

They allow for fast computation of determinants, powers, and inverses:

- The determinant is the product of its diagonal entries.
- A matrix power $\boldsymbol{D}^k$ is given by each diagonal element raised to the power k.
- The inverse $\boldsymbol{D}^{-1}$ is the reciprocal of its diagonal elements if all of them are nonzero.

Diagonalisable matrices and Diagonalisation
A matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is *diagonalisable* if it is similar to a diagonal matrix; meaning that there exists an invertible matrix $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ such that $\boldsymbol{D} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}$.

Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, let $\lambda_1, \ldots, \lambda_n$ be a set of scalars, and let $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n$ be a set of vectors in $\mathbb{R}^n$. We define $\boldsymbol{P} := [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n]$ and let $\boldsymbol{D} \in \mathbb{R}^{n\times n}$ be a diagonal matrix with diagonal entries $\lambda_1, \ldots, \lambda_n$. See that

$$\boldsymbol{A}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{D}$$

if and only if $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n$ are the corresponding eigenvectors of $\boldsymbol{A}$; since

$$\boldsymbol{A}\boldsymbol{P} = \boldsymbol{A}[\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n] = [\boldsymbol{A}\boldsymbol{p}_1, \ldots, \boldsymbol{A}\boldsymbol{p}_n]$$
$$\boldsymbol{P}\boldsymbol{D} = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n] \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} = [\lambda_1\boldsymbol{p}_1, \ldots, \lambda_n\boldsymbol{p}_n]$$

so

$$\boldsymbol{A}\boldsymbol{p}_1 = \lambda_1\boldsymbol{p}_1$$
$$\vdots$$
$$\boldsymbol{A}\boldsymbol{p}_n = \lambda_n\boldsymbol{p}_n$$

see that to obtain $\boldsymbol{D}$ we require $\boldsymbol{P}$ to be invertible, meaning $\boldsymbol{P}$ has full rank—this requires us to have n linearly independent eigenvectors $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n$.
(next page)

Eigendecomposition
Theorem: *A square matrix* $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ *can be factored into*

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$$

where $\boldsymbol{P} \in \mathbb{R}^{n \times n}$ *and* $\boldsymbol{D}$ *is a diagonal matrix whose diagonal entries are the eigenvalues of* $\boldsymbol{A}$*, if and only if the eigenvectors of* $\boldsymbol{A}$ *form a basis of* $\mathbb{R}$*.*

The requirement for n linearly independent eigenvectors means that only non-defective matrices can be diagonalised; due to the spectral theorem we have

Theorem: *A symmetric matrix* $\boldsymbol{S} \in \mathbb{R}^{n \times n}$ *can always be diagonalised.*

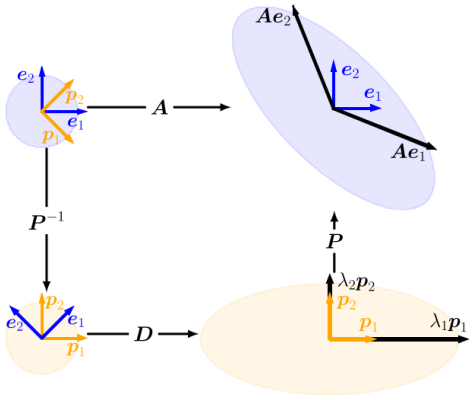
Moreover, the spectral theorem states that we can find an ONB basis of eigenvectors of $\mathbb{R}^n$; this means that $\boldsymbol{P}$ is an orthogonal matrix so that $\boldsymbol{D} = \boldsymbol{P}^T\boldsymbol{A}\boldsymbol{P}$.

A.3.11 More on Eigendecomposition

Geometric intuition for Eigendecomposition
The eigendecomposition formula is given as

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$$

We can interpret te eigendecomposition of a matrix as follows:



- First $\boldsymbol{P}^{-1}$ performs a basis change from the standard basis into the eigenbasis.
- Then, the diagonal matrix $\boldsymbol{D}$ scales the vectors along these axes by the eigenvalues λ_i.
- Finally, $\boldsymbol{P}$ transforms these scale vectors back into the standard/canonical coordinates.

Example

Consider computing the eigendecomposition of $\boldsymbol{A} = \frac{1}{2}\begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix}$

Step 1: Compute eigenvalues and eigenvectors
The characteristic polynomal of $\boldsymbol{A}$ is

$$\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = \det\left(\begin{bmatrix} \frac{5}{2} - \lambda & -1 \\ -1 & \frac{5}{2} - \lambda \end{bmatrix}\right) = \left(\lambda - \frac{7}{2}\right)\left(\lambda - \frac{3}{2}\right)$$

The eigenvalues of $\boldsymbol{A}$ are $\lambda_1 = \frac{7}{2}$ and $\lambda_2 = \frac{3}{2}$ (the roots of the characteristic polynomial). The associated normalised eigenvectors can then be obtained:

$$\boldsymbol{p}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \boldsymbol{p}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Note that because of the spectral theorem an orthonormal basis exists. (normalisation or projection methods may be required)
(next page)

Step 2: Check for existence
The eigenvectors $\boldsymbol{p}_1, \boldsymbol{p}_2$ form a basis of $\mathbb{R}^2$. Therefore, $\boldsymbol{A}$ can be diagonalised. (in this case the spectral theorem also gaurantees the existence of eigenvalues forming an ONB)

Step 3: Construct the matrix $\boldsymbol{P}$ to diagonalise $\boldsymbol{A}$
We collect the eigenvectors of $\boldsymbol{A}$ in $\boldsymbol{P}$:

$$\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2] = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

we also have our diagonal matrix where the diagonal elements are the eigenvalues:

$$\boldsymbol{D} = \begin{bmatrix} \frac{7}{2} & 0 \\ 0 & \frac{3}{2} \end{bmatrix}$$

We now have our eigendecomposition $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$:

$$\underbrace{\frac{1}{2}\begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix}}_{\boldsymbol{A}} = \underbrace{\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}}_{\boldsymbol{P}} \underbrace{\begin{bmatrix} \frac{7}{2} & 0 \\ 0 & \frac{3}{2} \end{bmatrix}}_{\boldsymbol{D}} \underbrace{\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}}_{\boldsymbol{P}^{-1}}$$

Significance of Eigendecomposition
The value of Eigendecompostion comes from the ease of computation of diagonal matrices:

- Diagonal matrices $\boldsymbol{D}$ can efficiently be raised to a power. This allows us to easily find the matrix power for a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ via eigendecomposition (since taking the power of a diagonal matrix is just the power of the diagonal elements):

  $$\boldsymbol{A}^k = (\boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1})^k = \boldsymbol{P}\boldsymbol{D}^k\boldsymbol{P}^{-1}$$

  Since computing $\boldsymbol{D}^k$ is efficient.

- Assuming that the eigendecomposition $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$ exists, then

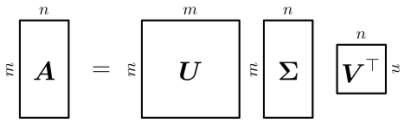
  $$\begin{aligned} \det(\boldsymbol{A}) &= \det(\boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}) = \det(\boldsymbol{P})\det(\boldsymbol{D})\det(\boldsymbol{P}^{-1}) \\ &= \det(\boldsymbol{D}) = \prod_i d_{ii} \end{aligned}$$

  Since $\det(\boldsymbol{P}^{-1}) = \frac{1}{\det(\boldsymbol{P})}$ and the determinant of a diagonal matrix is just the product down the diagonal.

Note that eigenvalue decomposition requires square matrices.

A.3.12 Singular Value Decomposition I

Theorem *Let* $\boldsymbol{A} =\in \mathbb{R}^{m\times n}$ *be a rectangular matrix of rank* $r \in [0, min(m, n)]$. *The SVD of* $\boldsymbol{A}$ *is the decomposition of the form*



$$\underbrace{\boldsymbol{A}}_{m\times n} = \underbrace{\boldsymbol{U}}_{m\times m}\underbrace{\boldsymbol{\Sigma}}_{m\times n}\underbrace{\boldsymbol{V}^T}_{n\times n}$$

- *with an orthogonal matrix* $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ *with column vectors* $\boldsymbol{u}_i, i = 1, \ldots, m$,
- *an orthogonal matrix* $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ *with column vectors* $\boldsymbol{v}_j, j = 1, \ldots, n$,
- *and a* $m \times n$ *matrix* $\boldsymbol{\Sigma}$ *with* $\Sigma_{ii} = \sigma_i \geq 0$ *and* $\Sigma_{ij} = 0, i \neq j$.

The diagonal entries $\sigma_i, i = 1, \ldots, r$, of $\boldsymbol{\Sigma}$ are called the *singular values*. The column vectors $\boldsymbol{u}_i$ are called the *left-singular vectors*. The column vectors $\boldsymbol{v}_j$ are called the *right-singular vectors*. (by convention the singular values are ordered, meaning $\sigma_1 \geq \sigma_2 \geq \sigma_r \geq 0$)
(next page)

Structure of the Singular value matrix $\boldsymbol{\Sigma}$

The *Singular value matrix* $\boldsymbol{\Sigma}$ is unique; observe that $\Sigma \in \mathbb{R}^{n\times n}$ is rectangular, being the same size as $\boldsymbol{A}$. This means that $\boldsymbol{\Sigma}$ has a diagonal submatrix that contains the singular values and needs additional zero padding outside the diagonal: if $m > n$, then the matrix $\boldsymbol{\Sigma}$ has diagonal structure up to row n and then consists of $\mathbf{0}^T$ row vectors from row $n+1$ to m below so that

$$\begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_n \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix}$$

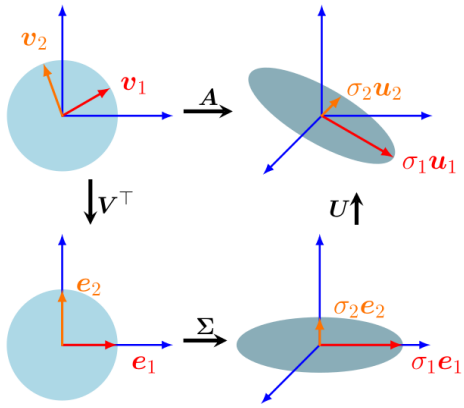
If $m < n$, the matrix $\boldsymbol{\Sigma}$ has diagonal structure up till column m and columns that consist of $\mathbf{0}$ from $m+1$ to n so that

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \ddots & 0 & \vdots & & \vdots \\ 0 & 0 & \sigma_m & 0 & \cdots & 0 \end{bmatrix}$$

The SVD exists for any matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$.

(next page)

Geometric intuition for the SVD
The SVD can be interpreted as a decomposition of a corresponding linear mapping $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ into three operations. Assume we are given a transformation matrix of a linear mapping $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ with respect to the standard bases B and C of $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively; moreover, assume a second basis $\tilde{B}$ of $\mathbb{R}^n$ and $\tilde{C}$ of $\mathbb{R}^m$. Then



1. The matrix $\boldsymbol{V}$ performs a basis change in the domain $\mathbb{R}^n$ from $\tilde{B}$ (represented by the red and orange vectors v_1 and v_2 in the top-left figure) to the standard basis B (since we set B to be the standard basis). $\boldsymbol{V}^T = \boldsymbol{V}^{-1}$ performs a basis change from B to $\tilde{B}$. (since $\boldsymbol{V}$ is defined as an orthogonal matrix). (The red and orange vectors are now aligned with the standard basis in the bottom left figure.)

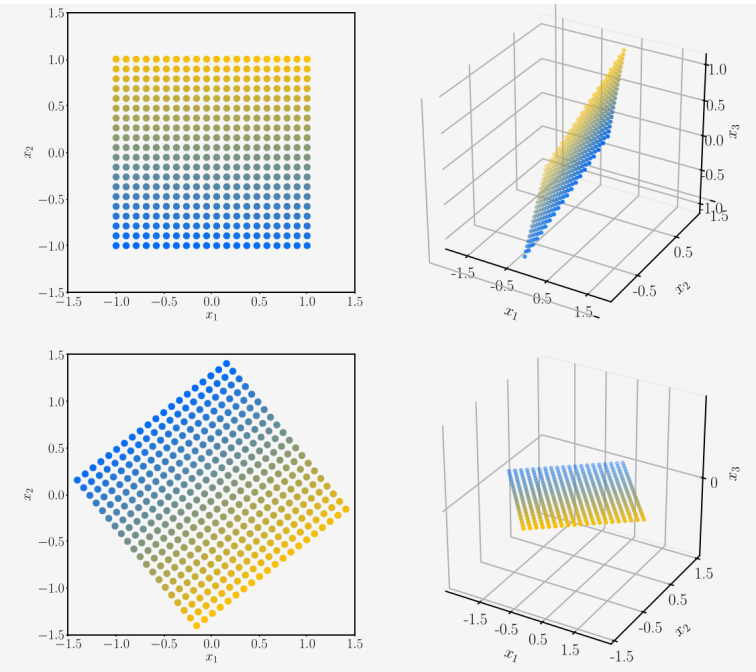
2. Having changed the coordinate system to $\tilde{B}$, $\boldsymbol{\Sigma}$ scales the new coordinates by the singular values σ_i (and adds or deletes dimensions); as with eigendecomposition, Σ can be viewed as the transformation matrix of Φ with respect to $\tilde{B}$ and $\tilde{C}$ (represented by the red and orange vectors being stretched, lying in the $\boldsymbol{e}_1\boldsymbol{e}_2$ plane, which is now embedded in a third dimension (bottom right figure).

3. $\boldsymbol{U}$ performs a basis change in the codomain $\mathbb{R}^m$ from $\tilde{C}$ into the basis of $\mathbb{R}^m$, represented by the rotation of the red and orange vectors out of the $\boldsymbol{e}_1\boldsymbol{e}_2$ plane (top right).

The SVD expresses a change of basis in both the domain and the codomain. This is in contrast with eigendecomposition which operates within the same vector space, where the same basis change is applied and then undone. What makes the SVD special is that these are two different bases linked by the singular value matrix Σ.
(next page)

Example
Consider the mapping of a square grid of vectors $\boldsymbol{\chi} \in \mathbb{R}^2$ that fit a box centered at the origin. Using the standard basis, we map these vectors using

$$\begin{aligned}\boldsymbol{A} &= \begin{bmatrix} 1 & -0.8 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \\ &= \begin{bmatrix} -0.79 & 0 & -0.62 \\ 0.38 & -0.78 & -0.49 \\ -0.48 & -0.62 & 0.62 \end{bmatrix} \begin{bmatrix} 1.62 & 0 \\ 0 & 1.0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} -0.78 & 0.62 \\ -0.62 & -0.78 \end{bmatrix}\end{aligned}$$



We start with a set of vectors $\boldsymbol{\chi}$ (top left, arranged in a grid). Applying $\boldsymbol{V}^T \in \mathbb{R}^{2\times 2}$ rotates $\boldsymbol{\chi}$ (bottom left). The singular value matrix $\boldsymbol{\Sigma}$ then maps these vectors to the codomain $\mathbb{R}^3$ (bottom right); note that up till now all these vectors lie in the x_1x_2 plane; the third coordinate is always zero, and the vectors in the x_1x_2 plane have been stretched by the singular values.

Finally $\boldsymbol{U}$ performs a rotation within the codomain $\mathbb{R}^3$; now the mapped vectors are no longer restricted to the x_1x_2 plane, but instead a different plane (top right).

A.3.13 Singular Value Decomposition II

SVD and Eigendecomposition
Compare the eigendecomposition of an SPD (symmetric positive definite) matrix

$$\boldsymbol{S} = \boldsymbol{S}^T = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^T$$

with the corresponding SVD

$$\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$$

see that if we set

$$\boldsymbol{U} = \boldsymbol{P} = \boldsymbol{V}, \quad \boldsymbol{D} = \boldsymbol{\Sigma}$$

the SVD of the SPD matrix is its eigendecomposition.

Construction of the SVD (intuition)
We begin with constructing the right-singular vectors. The spectral theorem tells us that the eigenvectors of a symmetric matrix form an ONB (which is also non-defective and therefore can undergo eigendecomposition). Since we can always construct a symmetric, positive semidefinite matrix $\boldsymbol{A}^T\boldsymbol{A} \in \mathbb{R}^{n\times n}$ from any rectangular matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, we can always diagonalise $\boldsymbol{A}^T\boldsymbol{A}$ and obtain

$$\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^T = \boldsymbol{P}\begin{bmatrix}\lambda_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \lambda_n\end{bmatrix}\boldsymbol{P}^T$$

Where $\boldsymbol{P}$ is an orthogonal matrix (because $\boldsymbol{A}^T\boldsymbol{A}$ is symmetric and has eigenvalues that can form an ONB). The $\lambda_i \geq 0$ are the eigenvalues of $\boldsymbol{A}^T\boldsymbol{A}$; assuming that the SVD of $\boldsymbol{A}$ exists:

$$\boldsymbol{A}^T\boldsymbol{A} = (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)^T(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T) = \boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$$

Where $\boldsymbol{U}, \boldsymbol{V}$ are orthogonal matrices (as defined in the SVD); therefore with $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ we obtain

$$\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}\boldsymbol{V}^T = \boldsymbol{V}\begin{bmatrix}\sigma_1^2 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sigma_n^2\end{bmatrix}\boldsymbol{V}^T$$

Comparing the two equations we get

$$\boldsymbol{V}^T = \boldsymbol{P}^T$$
$$\sigma_i^2 = \lambda_i$$

Therefore, the eigenvectors of $\boldsymbol{A}^T\boldsymbol{A}$ that compose $\boldsymbol{P}$ are the right-singular vectors $\boldsymbol{V}$ of $\boldsymbol{A}$; the eigenvalues of $\boldsymbol{A}^T\boldsymbol{A}$ are the squared singular values of $\boldsymbol{\Sigma}$.
(next page)

Construction of the SVD (cont.)
To obtain the left-singular vectors $\boldsymbol{U}$, we follow a similar procedure; notice that for a different symmetric matrix $\boldsymbol{A}\boldsymbol{A}^T \in \mathbb{R}^{m\times m}$ (instead of the previous $\boldsymbol{A}^T\boldsymbol{A} \in \mathbb{R}^{n\times n}$), we have

$$\begin{aligned}\boldsymbol{A}\boldsymbol{A}^T &= (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)^T = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T \\ &= \boldsymbol{U}\begin{bmatrix}\sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_m^2\end{bmatrix}\boldsymbol{U}^T\end{aligned}$$

since $\boldsymbol{A}\boldsymbol{A}^T$ is also a symmetric matrix (different from $\boldsymbol{A}^T\boldsymbol{A}$, notice they have different sizes for $m \neq n$), the spectral theorem tells us that as before, it can be diagonalised and an ONB of eigenvectors of $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{S}\boldsymbol{D}\boldsymbol{S}^T$ can be found and collected in $\boldsymbol{S}$; the orthonormal eigenvectors of $\boldsymbol{A}\boldsymbol{A}^T$ are the left-singular vectors $\boldsymbol{U}$ and form an orthonormal basis in the codomain of the SVD.

Also notice that $\boldsymbol{\Sigma}$ remains the same for both cases (since we use the SVD of $\boldsymbol{A}$), thus the nonzero entries σ_i in both cases are the same.

We finish the construction of the SVD by connecting the basis $\boldsymbol{U}$ to $\boldsymbol{V}$ using the fact that the images of the $\boldsymbol{v}_i$ under $\boldsymbol{A}$ have to be *orthogonal*. The inner product between $\boldsymbol{A}\boldsymbol{v}_i$ and $\boldsymbol{A}\boldsymbol{v}_j$ is 0 for $i \neq j$; for any two orthogonal eigenvectors $\boldsymbol{v}_i, \boldsymbol{v}_j$, $i \neq j$, see that

$$(\boldsymbol{A}\boldsymbol{v}_i)^T(\boldsymbol{A}\boldsymbol{v}_j) = \boldsymbol{v}_i^T(\boldsymbol{A}^T\boldsymbol{A})\boldsymbol{v}_j = \boldsymbol{v}_i^T(\lambda_j\boldsymbol{v}_j) = \lambda_j\boldsymbol{v}_i^T\boldsymbol{v}_j = 0$$

We need left-singular values that are *orthonormal*, so we normalise the images of the right-singular vectors $\boldsymbol{A}\boldsymbol{v}_i$ and obtain (since the images of $\boldsymbol{v}_i$ under $\boldsymbol{A}$ are orthogonal their normalisations are orthonormal)

$$\boldsymbol{u}_i := \frac{\boldsymbol{A}\boldsymbol{v}_i}{||\boldsymbol{A}\boldsymbol{v}_i||} = \frac{1}{\sqrt{\lambda_i}}\boldsymbol{A}\boldsymbol{v}_i = \frac{1}{\sigma_i}\boldsymbol{A}\boldsymbol{v}_i$$

the last equality comes from the earlier finding that $\sigma_i^2 = \lambda_i$. (point is that the normalisation factors turn out to be the singular values, connecting the orthonormal bases $\boldsymbol{U}$ and $\boldsymbol{V}$)
(next page)

Construction of the SVD (cont.)
We had

$$\boldsymbol{u}_i = \frac{1}{\sigma_i}\boldsymbol{A}\boldsymbol{v}_i$$

which can be rearranged into the *singular value equation*

$$\boldsymbol{A}\boldsymbol{v}_i = \sigma_i \boldsymbol{u}_i, \quad i = 1, \ldots, r$$

(this equation closely resembles the eigenvalue equation, but the vectors on the left and right hand sides are not the same)

Note that for $n < m$, this equation holds only for $i \leq n$, and says nothing about the $\boldsymbol{u}_i$ for $i > n$; (since $\boldsymbol{u}_i$ where $i > n$ get scaled to 0) however we know by construction that they are orthonormal. Conversely, for $m < n$, the equation holds only for $i \leq m$; for $i > m$, we have $\boldsymbol{A}\boldsymbol{v}_i = \boldsymbol{0}$ (as defined) since $\boldsymbol{v}_i$ form an orthonormal set; this means that the SVD also supplies an orthonormal basis of the kernel of $\boldsymbol{A}$.

A.3.14 Computing the SVD (Example)

Consider finding the singular value decomposition of

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{bmatrix}$$

the SVD requires us to compute the right-singular vectors $\boldsymbol{v}_j$, the singular values σ_k, and the left-singular vectors $\boldsymbol{u}_i$.

Step 1: Right-singular vectors as the eigenbasis of $\boldsymbol{A}^T\boldsymbol{A}$
We start by computing

$$\boldsymbol{A}^T\boldsymbol{A} = \begin{bmatrix} 1 & -2 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 5 & -2 & 1 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

Now we compute the singular values and right-singular values through eigendecomposition of $\boldsymbol{A}^T\boldsymbol{A}$, given as

$$\boldsymbol{A}^T\boldsymbol{A} = \begin{bmatrix} \frac{5}{\sqrt{30}} & 0 & \frac{-1}{\sqrt{6}} \\ \frac{-2}{\sqrt{30}} & \frac{1}{\sqrt{5}} & \frac{-2}{\sqrt{6}} \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{6}} \end{bmatrix} \begin{bmatrix} 6 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{5}{\sqrt{30}} & \frac{-2}{\sqrt{30}} & \frac{1}{\sqrt{30}} \\ 0 & \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{-1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{bmatrix} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^T$$

We obtain the right singular vectors as the columns of $\boldsymbol{P}$ so that

$$\boldsymbol{V} = \boldsymbol{P} = \begin{bmatrix} \frac{5}{\sqrt{30}} & 0 & \frac{-1}{\sqrt{6}} \\ \frac{-2}{\sqrt{30}} & \frac{1}{\sqrt{5}} & \frac{-2}{\sqrt{6}} \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{6}} \end{bmatrix}$$

Step 2: Singular value matrix
As the singular values σ_i are the square roots of the eigenvalues of $\boldsymbol{A}^T\boldsymbol{A}$ we obtain them straight from $\boldsymbol{D}$. Since $\text{rk}(\boldsymbol{A} = 2)$, there are only two nonzero singular values $\sigma_1 = \sqrt{6}$ and $\sigma_2 = 1$. The singular value matrix must be the same size as $\boldsymbol{A}$, and we obtain

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sqrt{6} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

(next page)

Step 3: Left-singular vectors as normalised image of right singular vectors

We find the left-singular vectors by computing the image of the right-singular vectors under $\boldsymbol{A}$ and normalising them by dividing them by their corresponding singular value; we obtain

$$\boldsymbol{u}_1 = \frac{1}{\sigma_1}\boldsymbol{A}\boldsymbol{v}_1 = \frac{1}{\sqrt{6}}\begin{bmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{bmatrix}\begin{bmatrix} \frac{5}{\sqrt{30}} \\ \frac{-2}{\sqrt{30}} \\ \frac{1}{\sqrt{30}} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{5}} \\ \frac{-2}{\sqrt{5}} \end{bmatrix}$$

$$\boldsymbol{u}_2 = \frac{1}{\sigma_2}\boldsymbol{A}\boldsymbol{v}_2 = \frac{1}{1}\begin{bmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{bmatrix}\begin{bmatrix} 0 \\ \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{bmatrix} = \begin{bmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{bmatrix}$$

$$\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2] = \frac{1}{\sqrt{5}}\begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix}$$

(On a computer the approach illustrated here has poor numerical behaviour, and the SVD of $\boldsymbol{A}$ is normally computed without resorting to eigenvalue decomposition of $\boldsymbol{A}^T\boldsymbol{A}$.)

A.3.15 $\boldsymbol{A^TA}$ and $\boldsymbol{AA^T}$ possess the same nonzero eigenvalues

Show that for any $\boldsymbol{A} \in \mathbb{R}^m \times n$ the matrices $\boldsymbol{A^TA}$ and $\boldsymbol{AA^T}$ possess the same nonzero eigenvalues. Consider an eigenvalue $\lambda \neq 0$ of $\boldsymbol{A^TA}$; meaning there exists $\boldsymbol{x} \in \mathbb{R}^n$ such that

$$\boldsymbol{A^TAx} = \lambda \boldsymbol{x}$$

now see that

$$\boldsymbol{AA^TAx} = \lambda \boldsymbol{Ax}$$

the eigenvector changed but the eigenvalue did not.

A.3.16 Eigenvalue Decomposition vs. Single Value Decomposition—Summary

Here we review the core elements of the eigendecomposition $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$ and the SVD $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$.

- The SVD always exists for any matrix $\mathbb{R}^{m\times n}$. The eigendecomposition is only defined for square matrices $\mathbb{R}^{n\times n}$ and only exists if we can find a basis of eigenvectors of $\mathbb{R}^n$.
- The vectors in the eigendecomposition matrix $\boldsymbol{P}$ are not necessarily orthogonal (the change of nasis is not a simple rotation and scaling). On the other hand, the vectors in the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ in the SVD are orthonormal, so they represent rotations.
- Both the eigendecomposition and the SVD are compositions of three linear mappings:
  1. Change of basis in the domain
  2. Independent scaling of each new basis vector and mapping from domain to codomain
  3. Change of basis in the codomain

  The difference here is that in the SVD, domain and codomain can be vector spaces of different dimensions.
- In the SVD, the left and right singular vector matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are generally not inverse of each other (since they perform basis changes in different vector spaces, and are of different sizes). In the eigendecomposition, the basis change matrices $\boldsymbol{P}$ and $\boldsymbol{P}^{-1}$ are inverses of each other.
- In the SVD, the entries in the diagonal matrix $\boldsymbol{\Sigma}$ are all real and nonnegative (due to the spectral theorem and the symmetry of $\boldsymbol{A}^T\boldsymbol{A}$), which is generally not true for the diagonal matrix in eigendecompostion.
- The SVD and the eigendecomposition are closely related through their projections:
  - The left singular vectors of $\boldsymbol{A}$ are the eigenvectors of $\boldsymbol{A}\boldsymbol{A}^T$
  - The right singular vectors of $\boldsymbol{A}$ are eigenvectors of $\boldsymbol{A}^T\boldsymbol{A}$
  - The nonzero singular values of $\boldsymbol{A}$ are the square roots of the nonzero eigenvalues of both $\boldsymbol{A}\boldsymbol{A}^T$ and $\boldsymbol{A}^T\boldsymbol{A}$.
- For symmetric matrices matrices $\boldsymbol{A} =\in \mathbb{R}^{n\times n}$, the eigenvalue decomposition and the SVD are one and the same, which follows from the spectral theorem.

A.3.17 Matrix Approximation

Intuituion
The SVD allows for factorisation of $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ into the product of three matrices

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$$

where $\boldsymbol{U} \in \mathbb{R}^{m \times m}$ and $\boldsymbol{V} \in \mathbb{R}^{n \times n}$ are orthogonal and $\boldsymbol{\Sigma}$ contains the singular values on its main diagonal.

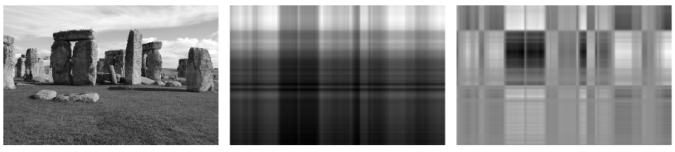
Rank-1 Matrices
We can construct a rank-1 matrix $\boldsymbol{A}_i \in \mathbb{R}^{m \times n}$ as

$$\boldsymbol{A}_i := \boldsymbol{u}_i \boldsymbol{v}_i^T$$

which is formed by the outer product of the ith orthogonal column vector of $\boldsymbol{U}$ and $\boldsymbol{V}$. Now see that, corresponding with the SVD, a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ of rank r can be written as a sum of rank-1 matrices $\boldsymbol{A}_i$ so that

$$\boldsymbol{A} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T = \sum_{i=1}^{r} \sigma_i \boldsymbol{A}_i$$

(visualise the SVD and notice that this is essentially the same) where the outer-product matrices $\boldsymbol{A}_i$ are weighted by the ith singular value σ_i; the diagonal structure of the singular value matrix $\boldsymbol{\Sigma}$ multiplies only matching left and right singular vectors $\boldsymbol{u}_i \boldsymbol{v}_i^T$ and scales them by their corresponding singular value σ_i. Illustrated:



(a) Original image $\boldsymbol{A}$. (b) $\boldsymbol{A}_1,\ \sigma_1 \approx 228,052$. (c) $\boldsymbol{A}_2,\ \sigma_2 \approx 40,647$.

(d) $\boldsymbol{A}_3,\ \sigma_3 \approx 26,125$. (e) $\boldsymbol{A}_4,\ \sigma_4 \approx 20,232$. (f) $\boldsymbol{A}_5,\ \sigma_5 \approx 15,436$.

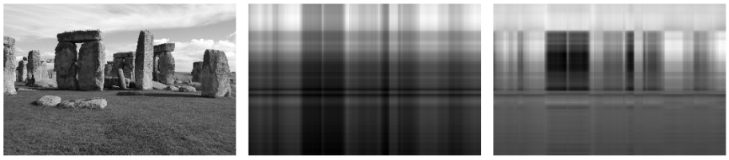
The original greyscale image is a 1432×1910 matrix of values between 0 (black) and 1 (white). (b) to (f) illustrate a few rank-1 matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_5$ and their corresponding singular values $\sigma_1, \ldots, \sigma_5$. Each matrix is formed by the outer-product of the left and right singular vectors.
(next page)

Approximation
Since summing up the r individual rank-1 matrices gives the rank-r matrix $\boldsymbol{A}$, if the sum dows not run over all matrices $\boldsymbol{A}_i, i = 1, \ldots, r$, but only up to an intermediate value $k < r$, we obtain a *rank-k approximation*

$$\widehat{\boldsymbol{A}}(k) := \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T = \sum_{i=1}^{k} \sigma_i \boldsymbol{A}_i$$

of $\boldsymbol{A}$ with $\mathrm{rk}(\widehat{\boldsymbol{A}}(k)) = k$.

Applied:



(a) Original image $\boldsymbol{A}$. (b) Rank-1 approximation $\widehat{\boldsymbol{A}}(1)$.(c) Rank-2 approximation $\widehat{\boldsymbol{A}}(2)$.

(d) Rank-3 approximation $\widehat{\boldsymbol{A}}(3)$.(e) Rank-4 approximation $\widehat{\boldsymbol{A}}(4)$.(f) Rank-5 approximation $\widehat{\boldsymbol{A}}(5)$.

Illustrated are low-rank approximations $\widehat{\boldsymbol{A}}(k)$ of an original image $\boldsymbol{A}$. The details of the image become increasingly recognisable in the rank-5 approximation. While the original image requires storage of $1432 \cdot 1910 = 2735120$ numbers, the rank-5 approximation only requires the storage of the five singular values and the five left and right singular vectors (1432×1 and 1×1910 each) for a total of $5 \cdot (1432 + 1910 + 1) = 16715$ numbers—about 0.6% of the original.

A.3.18 Spectral Norm, Eckart-Young theorem

Spectral Norm
For $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$, the *spectral norm* of a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is defined as

$$||\boldsymbol{A}||_2 := \max_{\boldsymbol{x}} \frac{||\boldsymbol{A}\boldsymbol{x}||_2}{||\boldsymbol{x}||_2}$$

The spectral norm essentially determines how long any vector $\boldsymbol{x}$ can at most become when multiplied by $\boldsymbol{A}$.

Spectral norm and Singular values
Theorem: *The spectral norm of $\boldsymbol{A}$ is its largest singular value σ_1.*

This can be seen from the SVD formula, where

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{x}$$

$\boldsymbol{U}$ and $\boldsymbol{V}^T$ are orthogonal matrices and therefore leave the norm unchanged. This leaves the 'scaling matrix' $\boldsymbol{\Sigma}$, where (by convention all $\sigma_i \geq 0$) the largest scaling factor is the largest singular value σ_1.

Eckart-Young Theorem
Theorem: *Consider a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ of rank r and let $\boldsymbol{B} \in \mathbb{R}^{m \times n}$ be a matrix of rank k. For any $k \leq r$ with $\widehat{\boldsymbol{A}}(k) = \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$ it holds that*

$$\widehat{\boldsymbol{A}}(k) = \text{argmin}_{\text{rk}(\boldsymbol{B})=k} ||\boldsymbol{A} - \boldsymbol{B}||_2$$
$$||\boldsymbol{A} - \widehat{\boldsymbol{A}}(k)||_2 = \sigma_{k+1}$$

The Eckart-Young theorem states explicitly how much error we introduce by approximating $\boldsymbol{A}$ using a rank-k approximation; we can interpret the rank-k approximation obtained with the SVD as a projection of the full rank matrix $\boldsymbol{A}$ onto a lower-dimensional space of rank-at-most-k matrices. Of all possible projections, the SVD minimises the error rate (with respect to the spectral norm) between $\boldsymbol{A}$ and any rank-k approximation.

Intuition for formula
Observe that the difference $\boldsymbol{A} - \widehat{\boldsymbol{A}}(k)$ is a matrix containing the sum of the remaining rank-1 matrices

$$\boldsymbol{A} - \widehat{\boldsymbol{A}}(k) = \sum_{i=k+1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$$

We immediately obtain σ_{k+1} as the spectral norm of the difference matrix.
(next page)

More intuition
Of all possible projections, the SVD minimises the error rate (with respect to the spectral norm) between $\boldsymbol{A}$ and any rank-k approximation. Say that this isn't the case and there is another matrix $\boldsymbol{B}$ with $\text{rk}(\boldsymbol{B}) \leq k$, such that

$$||\boldsymbol{A} - \boldsymbol{B}|| < ||\boldsymbol{A} - \widehat{\boldsymbol{A}}(k)||_2$$

(meaning that there exists an approximation with less error than the matrix approximation) then there exists an at least $(n-k)$-dimensional null space $Z \subseteq \mathbb{R}^n$, such that $\boldsymbol{x} \in Z$ implies that $\boldsymbol{B}\boldsymbol{x} = \boldsymbol{0}$. Then it follows that

$$||\boldsymbol{A}\boldsymbol{x}||_2 = ||(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{x}||_2$$

by using a version of the Cauchy-Schwartz inequality that encompasses norms of matrices, we obtain

$$||\boldsymbol{A}\boldsymbol{x}||_2 = \underbrace{||(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{x}||_2 < ||(\boldsymbol{A} - \boldsymbol{B})||_2 \, ||\boldsymbol{x}||_2}_{\text{Cauchy-Schwartz}} < \sigma_{k+1}||\boldsymbol{x}||_2$$

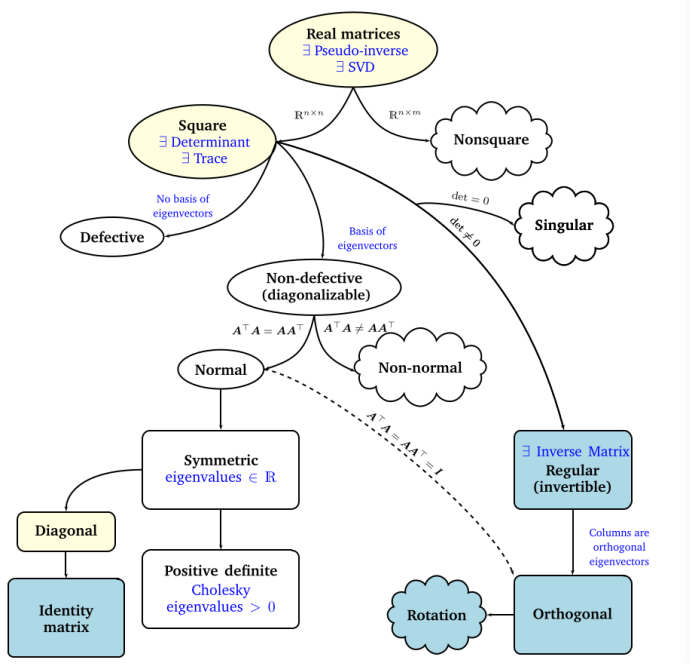
(the last inequality comes from our assumption that $||\boldsymbol{A} - \boldsymbol{B}|| < ||\boldsymbol{A} - \widehat{\boldsymbol{A}}(k)||_2 = \sigma_{k+1}$)

However, there exists a $(k+1)$-dimensional subspace where $||\boldsymbol{A}\boldsymbol{x}||_2 \geq \sigma_{k+1}||\boldsymbol{x}||_2$, which is spanned by the right-singular vectors $\boldsymbol{v}_j, j \leq k+1$ of $\boldsymbol{A}$. Adding up the dimensions of these two spaces yields a number greater than $\boldsymbol{n}$, which is a contradiction of the rank-nullity theorem.

Significance of Eckart-Young Theorem
The Eckart-Young Theorem implies that we can use the SVD to reduce a rank-r matrix $\boldsymbol{A}$ to a rank-k matrix in a principled, optimal manner. We can interpret the approximation of $\boldsymbol{A}$ by a rank-k matrix as a form of lossy compression.

A.3.19 Matrix Phylogeny (Summary of Chapters)

Illustrated is a fundamental phylogenic tree of matrices and linear mappings:



We consider all *real matrices* $\boldsymbol{A} \in \mathbb{R}^{n\times m}$; for non-square matrices the SVD always exists.

Square matrices

Focusing on *square matrices* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ the *determinant* tells one whether a square matrix possesses an *inverse matrix*. If the square $n \times n$ matrix possesses n linearly independent eigenvectors, then the matrix is non-defective and an *eigendecomposition* exists. We know that repeated eigenvalues may result in defective matrices, which cannot be diagonalised. (note that non-singular/non-invertible matrices and non-defective matrices are not the same)

(next page)

Normal, Orthogonal matrices
For non-defective square $n \times n$ matrices, $\boldsymbol{A}$ is *normal* if the contition $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^T$ holds. Moreover, if the more restrictive condition holds that $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I}$, then $\boldsymbol{A}$ is called *orthogonal*; the set of orthogonal matrices is a subset of the regular (invertible) matrices and satisfy $\boldsymbol{A}^T = \boldsymbol{A}^{-1}$.

Symmetric Matrices, Symmetric Positive definite Matrices
Normal matrices have a frequently encountered subset—the symmetric matrices $\boldsymbol{S} \in \mathbb{R}^n \times n$, which satisfy $\boldsymbol{S} = \boldsymbol{S}^T$. Symmetric matrices have only real eigenvalues (spectral theorem). A subset of the symmetric matrices consists of the positive definite matrices $\boldsymbol{P}$ which satisfy the condition of $\boldsymbol{x}^T\boldsymbol{P}\boldsymbol{x} > 0$ for all $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$; in this case, a unique *Cholesky decomposition* exists. Positive definite matrices have only positive eigenvalues and are always invertible.

Diagonal matrices
Another subset of symmetric matrices consists of the *diagonal matrices* $\boldsymbol{D}$; diagonal matrices are closed under multiplication and addition, but do not necessarily form a group. A special diagonal matrix is the identity $\boldsymbol{I}$.

A.4 Vector Calculus

A.4.1 Partial Differentiation and Gradients

Definition
The generalisaion of the derivative to functions of several variables is the *gradient*; the gradient of function f with respect to x is found by *varying one variable at a time* and keeping others constant. The gradient is then the collection of these *partial derivatives*.

For a function $f : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{x} \mapsto f(x)$, $\boldsymbol{x} \in \mathbb{R}^n$ we define the *partial derivatives* as

$$\begin{aligned}\frac{\partial f}{\partial x_1} &= \lim_{h\to 0} \frac{f(x_1 + h, x_2, \ldots, x_n) - f(\boldsymbol{x})}{h} \\ &\vdots \\ \frac{\partial f}{\partial x_n} &= \lim_{h\to 0} \frac{f(x_1, \ldots, x_{n-1}, x_n + h) - f(\boldsymbol{x})}{h}\end{aligned}$$

collecting them in a row vector we get

$$\nabla_x f = \text{grad} f = \frac{df}{d\boldsymbol{x}} = \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} & \frac{\partial f(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{1\times n}$$

this is called the *gradient* of f (it is also the *Jacobian*, but note that this is a particular case of the Jacobian, whose definition can apply more generally to vector-valued functions).

A.4.2 Chain Rule

Definition and Vector notation
Considering a function $f : \mathbb{R}^2 \to \mathbb{R}$ of two variables $x_1 x_2$, where $x_1(t)$ and $x_2(t)$ are themselves functions of t, to compute the gradient of f with respect to t we apply the chain rule for multivariate functions as

$$\frac{df}{dt} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1(t)}{\partial t} \\ \frac{\partial x_2(t)}{\partial t} \end{bmatrix} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t}$$

Consider another case where $f(x_1, x_2)$ is a function of x_1 and x_2, where $x_1(s,t)$ and $x_2(s,t)$ are themselves functions of two variables s and t, the chain rule yields

$$\frac{df}{ds} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial s} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial s}$$
$$\frac{df}{dt} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t}$$

expressed compactly in matrix notation we have

$$\frac{df}{d(s,t)} = \frac{\partial f}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial (s,t)} = \underbrace{\begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix}}_{=\frac{\partial f}{\partial \boldsymbol{x}}} \underbrace{\begin{bmatrix} \frac{\partial x_1}{\partial s} & \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial s} & \frac{\partial x_2}{\partial t} \end{bmatrix}}_{=\frac{\partial \boldsymbol{x}}{\partial (s,t)}}$$

A.4.3 Gradients of Vector-Valued Functions, the Jacobian

Intuition

Here we generalise the concept of the gradient to vector valued functions $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$, where $n \geq 1$ and $m > 1$. Considering a function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ and a vector $\boldsymbol{x} = [x_1, \ldots, x_n]^T \in \mathbb{R}^n$, the corresponding vector of function values is given as

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_m(\boldsymbol{x}) \end{bmatrix} \in \mathbb{R}^m$$

Writing the vector-valued function in this way allows us to view a vector-valued function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ as a vector of functions $[f_1, \ldots, f_m]^T$, $f_i : \mathbb{R}^n \to \mathbb{R}$ that map onto $\mathbb{R}$.

Therefore, the partial derivative of a vector-valued function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ with respect to $x_i \in \mathbb{R}$, $i = 1, \ldots, n$, is given as the vector

$$\frac{\partial \boldsymbol{f}}{\partial x_i} = \begin{bmatrix} \frac{\partial f_1}{\partial x_i} \\ \vdots \\ \frac{\partial f_m}{\partial x_i} \end{bmatrix} = \begin{bmatrix} \lim_{h \to 0} \frac{f_1(x_1, \ldots, x_{i-1}, x_i + h, x_{i+1}, \ldots, x_n) - f_1(\boldsymbol{x})}{h} \\ \vdots \\ \lim_{h \to 0} \frac{f_m(x_1, \ldots, x_{i-1}, x_i + h, x_{i+1}, \ldots, x_n) - f_m(\boldsymbol{x})}{h} \end{bmatrix} \in \mathbb{R}^m$$

The gradient of $\boldsymbol{f}$ with respect to a vector is the row vector of the partial derivatives, and every partial derivative $\partial \boldsymbol{f} / \partial x_i$ is itself a column vector. Therefore we obtain the gradient of $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ with respect to $\boldsymbol{x} \in \mathbb{R}^n$ by collecting these partial derivatives:

$$\begin{aligned} \frac{d\boldsymbol{f}(x)}{d\boldsymbol{x}} &= \begin{bmatrix} \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n} \end{aligned}$$

(next page)

The Jacobian
The collection of all first-order partial derivatives of a vector-valued function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ is called the *Jacobian*. The Jacobian $\boldsymbol{J}$ is an $m \times n$ matrix, which we define and arrange as follows:

$$\begin{aligned}\boldsymbol{J} = \nabla_{\boldsymbol{x}} \boldsymbol{f} = \frac{d\boldsymbol{f}(x)}{d\boldsymbol{x}} &= \begin{bmatrix} \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \\ \text{where } \boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad J(i,j) &= \frac{\partial f_i}{\partial x_i}\end{aligned}$$

(see that a particular case of the Jacobian is that for a function $f : \mathbb{R}^n \to \mathbb{R}^1$, which possesses a Jacobian that is a row vector)

(Also note that the above notation of the Jacobian is termed the *numerator layout* of the derivative, where the elements of $\boldsymbol{f}$ define the rows while the elements of $\boldsymbol{x}$ define the columns. There exists also a *denominator layout*, which is the transpose of the numerator layout.)

Approaches to identifying basis change matrices
Consider a basis change from $(\boldsymbol{b}_1, \boldsymbol{b}_2)$ to $(\boldsymbol{c}_1, \boldsymbol{c}_2)$, say $\boldsymbol{b}_1 = [1, 0]^T, \boldsymbol{b}_2 = [0, 1]^T$ and $\boldsymbol{c}_1 = [-2, 1]^T, \boldsymbol{c}_2 = [1, 1]^T$, by expressing the new basis in terms of the old basis we can compute the basis change matrix. Here we show (intuitively) that partial derivatives can also provide a general approach to finding such mappings.

Approach 1: Using the aforementioned method we can identify the desired basis change matrix as

$$\boldsymbol{J} = \begin{bmatrix} -2 & 1 \\ 1 & 1 \end{bmatrix}$$

such that $\boldsymbol{J}\boldsymbol{b}_1 = \boldsymbol{c}_1$ and $\boldsymbol{J}\boldsymbol{b}_2 = \boldsymbol{c}_2$. (The mapping has a determinant of absolute value 3; in this case, since the old basis cover a square of area 1, we can conclude that the area spanned by the new basis is three times greater than that of the old basis)
(next page)

Approach 2: Consider a function $\boldsymbol{f} : \mathbb{R}^2 \to \mathbb{R}^2$ that performs a variable transformation; in this case it maps the coordinate representation pf any vector $\boldsymbol{x} \in \mathbb{R}^2$ with respect to $(\boldsymbol{b}_1, \boldsymbol{b}_2)$ to the coordinate representaton $\boldsymbol{y}$ with respect to $(\boldsymbol{c}_1, \boldsymbol{c}_2)$.

We want to identify themapping so that we can compute how an area/volume changes under transformation by $\boldsymbol{f}$. For this we need to find out how $\boldsymbol{f}(\boldsymbol{x})$ changes under small changes to $\boldsymbol{x}$, leading us to take partial derivatives; since we can write

$$\begin{aligned} y_1 &= -2x_1 + x_2 \\ y_2 &= x_1 + x_2 \end{aligned}$$

we obtain the partial derivatives—the functional relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$\frac{\partial y_1}{\partial x_1} = -2, \quad \frac{\partial y_1}{\partial x_2} = 1, \quad \frac{\partial y_2}{\partial x_1} = 1, \quad \frac{\partial y_2}{\partial x_2} = 1$$

and compose the Jacobian as

$$\boldsymbol{J} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ 1 & 1 \end{bmatrix}$$

see that the Jacobian represents the coordinate transformation exactly if the transformation is linear (as in this case), and thus recovers the basis change matrix. Also notice therefore the significance of the *Jacobian determinant* $|\det(\boldsymbol{J})|$ (as in the first approach).

A.4.4 Gradient of Least-Squares Loss in a Linear Model

Consider the linear model

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{\theta}$$

where $\boldsymbol{\theta} \in \mathbb{R}^D$ is a parameter vector, $\Phi \in \mathbb{R}^{N \times D}$ are input features and $\boldsymbol{y} \in \mathbb{R}^N$ are the corresponding observations. We define the functions

$$L(e) := ||\boldsymbol{e}||^2$$
$$\boldsymbol{e}(\boldsymbol{\theta}) := \boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}$$

(We want to optimise for minimal loss) L is called a *least squares loss* function; we seek $\partial L/\partial \boldsymbol{\theta}$. First we determine its dimensionality:

$$\frac{\partial L}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{1 \times D}$$

using the chain rule we can compute the gradient as

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = \underbrace{\frac{\partial L}{\partial \boldsymbol{e}}}_{1 \times N} \underbrace{\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}}}_{N \times D}$$

where the dth element is given by

$$\frac{\partial L}{\partial \boldsymbol{\theta}}[1, d] = \sum_{n=1}^{N} \frac{\partial L}{\partial \boldsymbol{e}}[n] \frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}}[n, d]$$

assuming the euclidean norm we have the function $L = ||\boldsymbol{e}||^2 = \boldsymbol{e}^T\boldsymbol{e}$; we can therefore determine

$$\frac{\partial L}{\partial \boldsymbol{e}} = 2\boldsymbol{e}^T \in \mathbb{R}^{1 \times N}$$

from the second equation we can also obtain

$$\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}} = -\boldsymbol{\Phi} \in \mathbb{R}^{N \times D}$$

With that our desired derivative is

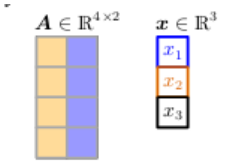
$$\frac{\partial L}{\partial \boldsymbol{\theta}} = -2\boldsymbol{e}^T\boldsymbol{\Phi} = -2(\boldsymbol{y}^T - \boldsymbol{\theta}^T\boldsymbol{\Phi}^T)\Phi \in \mathbb{R}^{1 \times D}$$

A.4.5 Gradients of Matrices, Tensors

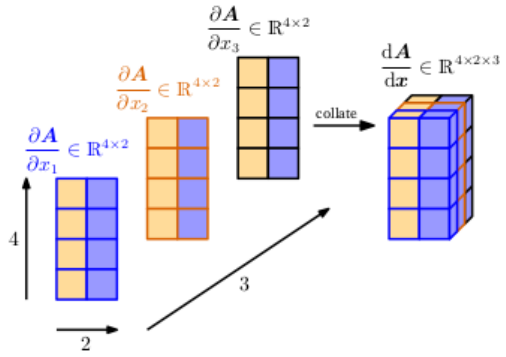
Should we require the gradient of matrices with respect to vectors (or other matrices), this leads to a multidimensional tensor; we can think of a tensor as a multidimensional array that collects partial derivatives.

For instance consider computing the gradient of an $m \times n$ matrix $\boldsymbol{A}$ with respect to a $p \times q$ matrix $\boldsymbol{B}$; the resulting jacobian would have dimension $(m \times n) \times (p \times q)$ (a four dimensional tensor), whose entries are given as $J_{ijkl} = \partial A_{ij} / \partial B_{kj}$.

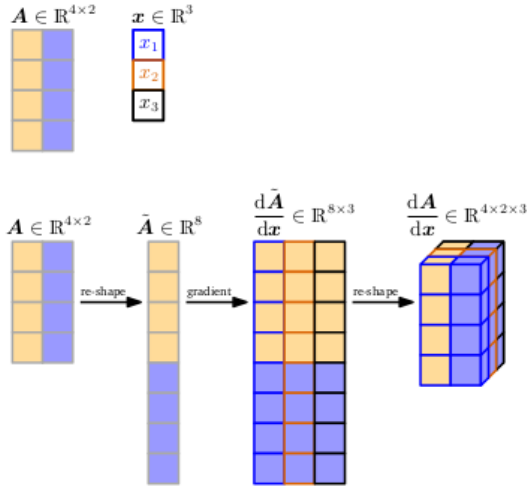
Notice, however, that since matrices represent linear mappings, we can exploit the fact that there is a vector-space isomorphism (a linear, invertible mapping) between the space $\mathbb{R}^{m \times n}$ and $\mathbb{R}^{mn}$—we can reshape the matrices into vectors of lengths mn and pq respectively to obtain a jacobian of $mn \times pq$ (illustrated on the next page).
(next page)



Partial derivatives:



(a) Approach 1: We compute the partial derivative $\frac{\partial \boldsymbol{A}}{\partial x_1}, \frac{\partial \boldsymbol{A}}{\partial x_2}, \frac{\partial \boldsymbol{A}}{\partial x_3}$, each of which is a 4×2 matrix, and collate them in a $4 \times 2 \times 3$ tensor.



(b) Approach 2: We re-shape (flatten) $\boldsymbol{A} \in \mathbb{R}^{4\times 2}$ into a vector $\tilde{\boldsymbol{A}} \in \mathbb{R}^8$. Then, we compute the gradient $\frac{\mathrm{d}\tilde{\boldsymbol{A}}}{\mathrm{d}\boldsymbol{x}} \in \mathbb{R}^{8\times 3}$. We obtain the gradient tensor by re-shaping this gradient as illustrated above.

A.4.6 Gradients of Matrices—Examples

Example 1—Vector and Matrix
Given

$$\boldsymbol{f} = \boldsymbol{A}\boldsymbol{x}, \quad \boldsymbol{f} \in \mathbb{R}^M, \quad \boldsymbol{A} \in \mathbb{R}^{M \times N}, \quad \boldsymbol{x} \in \mathbb{R}^N$$

should we want to find $d\boldsymbol{f}/d\boldsymbol{A}$ the dimension of our gradient would be

$$\frac{d\boldsymbol{f}}{d\boldsymbol{A}} \in \mathbb{R}^{M \times (M \times N)}$$

By definition, the gradient is a collection of partial derivatives, where

$$\frac{d\boldsymbol{f}}{d\boldsymbol{A}} = \begin{bmatrix} \frac{\partial f_1}{\partial \boldsymbol{A}} \\ \vdots \\ \frac{\partial f_M}{\partial \boldsymbol{A}} \end{bmatrix}, \quad \frac{\partial f_i}{\partial \boldsymbol{A}} \in \mathbb{R}^{1 \times (M \times N)}$$

To compute the partial derivatives (we require expressions for each function) it helps to explicitly write out the matrix vector multiplication:

$$f_i = \sum_{j=1}^{N} A_{ij} x_j, \quad i = 1, \ldots, M$$

with that the partial derivatives are given as

$$\frac{\partial f_i}{\partial A_{iq}} = x_q$$

written with respect to a row:

$$\frac{\partial f_i}{\partial A_{i,:}} = \boldsymbol{x}^T \in \mathbb{R}^{1 \times 1 \times N}$$
$$\frac{\partial f_i}{\partial A_{k \neq i,:}} = \boldsymbol{0}^T \in \mathbb{R}^{1 \times 1 \times N}$$

where we obtain a $1 \times 1 \times N$ sized tensor as the partial derivative of f_i with respect to a row of $\boldsymbol{A}$. Stacking the partial derivatives of the rows we get the desired gradient

$$\frac{\partial f_i}{\partial \boldsymbol{A}} = \begin{bmatrix} \boldsymbol{0}^T \\ \vdots \\ \boldsymbol{0}^T \\ \boldsymbol{x}^T \\ \boldsymbol{0}^T \\ \vdots \\ \boldsymbol{0}^T \end{bmatrix} \in \mathbb{R}^{1 \times (M \times N)}$$

(next page)

Example 2—Matrix and Matrix
Considering a matrix $\boldsymbol{R} \in \mathbb{R}^{M\times N}$ and $\boldsymbol{f} : \mathbb{R}^{M\times N} \to \mathbb{R}^{N\times N}$ with

$$\boldsymbol{f}(\boldsymbol{R}) = \boldsymbol{R}^T\boldsymbol{R} =: \boldsymbol{K} \in \mathbb{R}^{N\times N}$$

where we seek the gradient $d\boldsymbol{K}/d\boldsymbol{R}$.

The gradient has dimensionality

$$\frac{d\boldsymbol{K}}{d\boldsymbol{R}} \in \mathbb{R}^{(N\times N)\times(M\times N)}$$

which is a tensor, where

$$\frac{dK_{pq}}{d\boldsymbol{R}} \in \mathbb{R}^{1\times(M\times N)}$$

for $p, q = 1, \ldots, N$; where K_{pq} is the (p, q)th entry of $\boldsymbol{K} = \boldsymbol{f}(\boldsymbol{R})$, denoting the ith column of $\boldsymbol{R}$ by $\boldsymbol{r}_i$, every entry of $\boldsymbol{K}$ is given by the dot product of two columns of $\boldsymbol{R}$:

$$K_{pq} = \boldsymbol{r}_p^T\boldsymbol{r}_q = \sum_{m=1}^{M} R_{mp}R_{mq}$$

now we compute the partial derivative $\frac{\partial K_{pq}}{\partial R_{ij}}$ and obtain

$$\frac{\partial K_{pq}}{\partial R_{ij}} = \sum_{m=1}^{M} \frac{\partial}{\partial R_{ij}} R_{mp}R_{mq} = \partial_{pqij}$$

where

$$\partial_{pqij} = \begin{cases} R_{iq} & \text{if } j = p, p \neq q \\ R_{ip} & \text{if } j = q, p \neq q \\ 2R_{iq} & \text{if } j = p, p = q \\ 0 & \text{otherwise} \end{cases}$$

the desired gradient has the dimension $(N \times N) \times (M \times N)$, and every single entry of this tensor is given by ∂_{pqij}, where $p, q, j = 1, \ldots, N$ and $i = 1, \ldots, M$.

A.4.7 Backpropagation

Motivation
Given a function

$$f(x) = \sqrt{x^2 + \exp(x^2)} + \cos(x^2 + \exp(x^2))$$

By application of the chain rule, and noting that differentiation is linear, we can compute the gradient as

$$\frac{df}{dx} = 2x \left(\frac{1}{2\sqrt{x^2 + \exp(x^2)}} - \sin(x^2 + \exp(x^2)) \right) (1 + \exp(x^2))$$

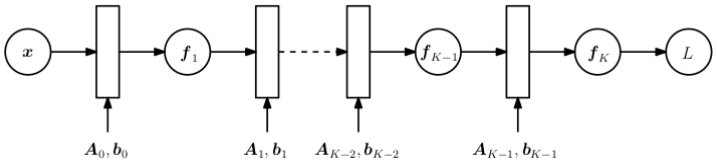
Writing out the gradient in this explicit way is often impractical, and could be computationally expensive. When training neural network models, the *back-propagation algorithm* is an effivient way to compute the gradient of an error function with respect to parameters we want to optimise over.

Backpropagation
In deep learning, a function value $\boldsymbol{y}$ is computed as a many-level function composition:

$$\boldsymbol{y} = (f_K \circ f_{K-1} \circ \cdots \circ f_1)(\boldsymbol{x}) = f_K(f_{K-1}(\cdots(f_1(\boldsymbol{x}))\cdots))$$

where $\boldsymbol{x}$ are the inputs and $\boldsymbol{y}$ are the observations, with every function f_i possessing its own parameters.

The sequential application of such functions are organised into multiple layers:



we have functions $f_i(x_{i-1}) = \sigma(\boldsymbol{A}_{i-1}\boldsymbol{x}_{i-1} + \boldsymbol{b}_{i-1})$ in the ith layer, where $\boldsymbol{x}_{i-1}$ is the output of the layer $i-1$ and σ an activation function like the logistic sigmoid $\frac{1}{1+e^{-x}}$. To train these models we require a gradient of the loss function with respect to all model parameters $\boldsymbol{A}_j, \boldsymbol{b}_j$ where $j = 1, \ldots, K$; thus we need to compute the gradients with respect to each layer.

(next page)

(cont.) Say we have inputs $\boldsymbol{x}$ and observations $\boldsymbol{y}$ and a network structure defined by

$$\begin{aligned}\boldsymbol{f}_0 &:= \boldsymbol{x} \\ \boldsymbol{f}_i &:= \sigma_i(\boldsymbol{A}_{i-1}\boldsymbol{f}_{i-1} + \boldsymbol{b}_{i-1}), \quad i = 1, \ldots, K\end{aligned}$$

we may be interested in finding $\boldsymbol{A}_j, \boldsymbol{b}_j$ for $j = 0, \ldots, K-1$ such that the squared loss

$$L(\boldsymbol{\theta}) = ||\boldsymbol{y} - \boldsymbol{f}_K(\boldsymbol{\theta}, \boldsymbol{x})||^2$$

is minimised, where $\boldsymbol{\theta} = \{\boldsymbol{A}_0, \boldsymbol{b}_0, \ldots, \boldsymbol{A}_{K-1}, \boldsymbol{b}_{K-1}\}$.

To obtain the gradients with respect to the parameter set $\boldsymbol{\theta}$, we require the partial derivatives of L with respect to the parameters $\boldsymbol{\theta}_j = \{\boldsymbol{A}_j, \boldsymbol{b}_j\}$ of each layer $j = 0, \ldots, K-1$. The chain rule allows us to determine the partial derivatives as

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{\theta}_{K-1}} &= \frac{\partial L}{\partial \boldsymbol{f}_K}\frac{\partial \boldsymbol{f}_K}{\partial \boldsymbol{\theta}_{K-1}} \\ \frac{\partial L}{\partial \boldsymbol{\theta}_{K-2}} &= \frac{\partial L}{\partial \boldsymbol{f}_K}\frac{\partial \boldsymbol{f}_K}{\partial \boldsymbol{f}_{K-1}}\frac{\partial \boldsymbol{f}_{K-1}}{\partial \boldsymbol{\theta}_{K-2}} \\ \frac{\partial L}{\partial \boldsymbol{\theta}_{K-3}} &= \frac{\partial L}{\partial \boldsymbol{f}_K}\frac{\partial \boldsymbol{f}_K}{\partial \boldsymbol{f}_{K-1}}\frac{\partial \boldsymbol{f}_{K-1}}{\partial \boldsymbol{f}_{K-2}}\frac{\partial \boldsymbol{f}_{K-2}}{\partial \boldsymbol{\theta}_{K-3}} \\ \frac{\partial L}{\partial \boldsymbol{\theta}_i} &= \frac{\partial L}{\partial \boldsymbol{f}_K}\frac{\partial \boldsymbol{f}_K}{\partial \boldsymbol{f}_{K-1}} \cdots \frac{\partial \boldsymbol{f}_{i+2}}{\partial \boldsymbol{f}_{i+1}}\frac{\partial \boldsymbol{f}_{i+1}}{\partial \boldsymbol{\theta}_i}\end{aligned}$$

see that, assuming one has already computed the partial derivatives $\partial L/\partial \boldsymbol{\theta}_{i+1}$, then most of the computation can be reused to compute $\partial L/\partial \boldsymbol{\theta}_i$.

A.4.8 Automatic Differentiation

Backpropagation is really a special case of a general technique of numerical analysis called *automatic differentiaton*. We can think of automatic differentiation as a set of techniques to numerically evaluate the gradient of a function by working with intermediate variables and applying the chain rule.

Consider data flow from inputs x to y via some intermediate variables a, b; if we were to compute the derivative dy/dx, we would apply the chain rule and obtain

$$\frac{dy}{dx} = \frac{dy}{db}\frac{db}{da}\frac{da}{dx}$$

Also see that due to the associativity of matrix multiplication, we can choose between *forward mode* and *reverse mode*:

$$\frac{dy}{dx} = \left(\frac{dy}{db}\frac{db}{da}\right)\frac{da}{dx}$$
$$\frac{dy}{dx} = \frac{dy}{db}\left(\frac{db}{da}\frac{da}{dx}\right)$$

Where the first equation would be reverse mode since the gradients are propagated backward through the graph; while the second being forward mode.

Formalisation
Let $x_1, \ldots, x_d$ be the input variables to a function, $x_{d+1}, \ldots, x_{D-1}$ be the intermediate variables, and x_D the output variable; then the computation graph can be expressed as follows

$$\text{For } i = d+1, \ldots, D: \quad x_i = g_i(x_{\text{Pa}(x_i)})$$

where the $g_i(\cdot)$ are elementary functions and $x_{\text{Pa}(x_i)}$ are the parent nodes of the variable x_i in the computation graph. Given a function defined this way we can use the chain rule to compute the derivative of the function in a step-by-step fashion. We have $f = x_D$ (as defined since f is the output); for all other variables we apply the chain rule:

$$\frac{\partial f}{\partial x_i} = \sum_{x_j : x_i \in \text{Pa}(x_j)} \frac{\partial f}{\partial x_j}\frac{\partial x_j}{\partial x_i} = \sum_{x_j : x_i \in \text{Pa}(x_j)} \frac{\partial f}{\partial x_j}\frac{\partial g_j}{\partial x_i}$$

The automatic differentiation approach works whenever we have a function that can be expressed as a computation graph with differentiable elementary functions.
(next page)

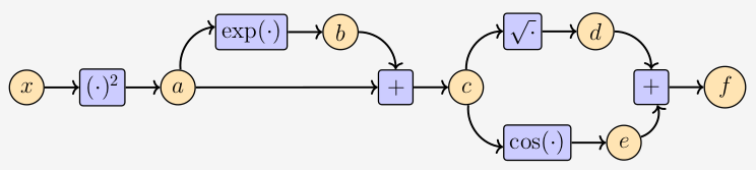
Example
Automatic differentiation is essentially the formalisation of the following instructive example. Here we use reverse mode automatic differentiation (which in the context of neural networks is computationally significantly cheaper due to the tendency for high input dimensionality): Consider the function

$$f(x) = \sqrt{x^2 + \exp(x^2)} + \cos(x^2 + \exp(x^2))$$

see that if we were to implement a function f on a computer, we would be able to save some computation by using *intermediate variables*:

$$\begin{aligned} a &= x^2, & b &= \exp(a) \\ c &= a + b, & d &= \sqrt{c} \\ e &= \cos(c), & f &= d + e \end{aligned}$$

Notice how the last variable is 'connected' to the first variable by a sequence of intermediate variables (this is the same kind of thinking that occurs when applying the chain rule). Also note that the preceding set of equations requires fewer operations than a direct implementation of the function as fully defined in the beginning; the corresponding *computation graph* shows the flow of data and computations required to obtain the final function output:



computation graphs are representations that are widely used in implementation of neural network software libraries. We can directly compute the derivatives of the intermediate variables with respect to their corresponding inputs (which is easy since they are elementary functions):

$$\begin{aligned} \frac{\partial a}{\partial x} &= 2x, & \frac{\partial b}{\partial a} &= \exp(a) \\ \frac{\partial c}{\partial a} = 1 = \frac{\partial c}{\partial b}, & & \frac{\partial d}{\partial c} &= \frac{1}{2\sqrt{c}} \\ \frac{\partial e}{\partial c} &= -\sin(c), & \frac{\partial f}{\partial d} &= 1 = \frac{\partial f}{\partial e} \end{aligned}$$

(next page)

and by looking at the computation graph we can compute $\partial f/\partial x$ by working backward from the output to obtain

$$\begin{aligned}
\frac{\partial f}{\partial c} &= \frac{\partial f}{\partial d}\frac{\partial d}{\partial c} + \frac{\partial f}{\partial e}\frac{\partial e}{\partial c} \\
\frac{\partial f}{\partial b} &= \frac{\partial f}{\partial c}\frac{\partial c}{\partial b} \\
\frac{\partial f}{\partial a} &= \frac{\partial f}{\partial b}\frac{\partial b}{\partial a} + \frac{\partial f}{\partial c}\frac{\partial c}{\partial a} \\
\frac{\partial f}{\partial x} &= \frac{\partial f}{\partial a}\frac{\partial a}{\partial x}
\end{aligned}$$

where we think of each of the derivatives as a variable.

A.4.9