SSIE Department *Binghamton University*

Notes on Linear Algebra

dcline1@binghamton.edu *Daniel A. Cline*

These notes are primarily based on Strang [4] and Strang [5] and are intended as a quick reference for more advanced study in machine learning (e.g. Hastie at. al. [2], Murphy [3]) and convex optimization (e.g. Boyd & Vandenberghe [1]).[1]

1 Vectors

By convention, vectors are taken to be **column vector**s in most scientific writing, and are often written as $v = (v_1, \dots, v_n)$ or $v = [v_1 \dots v_n]^T$ to save space, where v is an n-dimensional column vector ($v \in \mathbf{R}^{n \times 1}$, or simply $v \in \mathbf{R}^n$).

The two most fundamental operations that may be performed on vectors are **scalar multiplication** (cv, where $c \in \mathbf{R}$ is a real number) and **vector addition** ($v + w$, where $v \in \mathbf{R}^n$ and $w \in \mathbf{R}^n$). When both operations are combined, we get a **linear combination** of vectors ($cv + dw$).

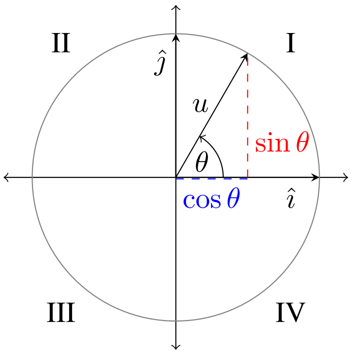
The **dot product** (inner product) of vectors $v \in \mathbf{R}^n$ and $w \in \mathbf{R}^n$ is given by

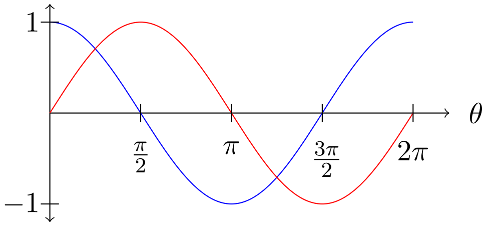
$$v \cdot w = v^T w = \sum_{i=1}^{n} v_i w_i \quad \text{(commutative: } v \cdot w = w \cdot v\text{).}$$

The **length** (Euclidean norm, 2-norm) of a vector v is given by $\|v\| = \sqrt{v_1^2 + \cdots + v_n^2} = \sqrt{v \cdot v}$. A **unit vector** u is a vector with length 1 ($\|u\| = 1$). For any vector v, the vector $u = v/\|v\|$ is a unit vector that points in the same direction as v.

Vectors are **orthogonal** (perpendicular, the angle between them is 90°) if their dot product is zero. Unit vectors are **orthonormal** (often denoted q_i) if they are orthogonal. A simple example is the **standard basis** vectors, which are of length one and are always orthogonal. The standard basis vectors in $\mathbf{R}^2$ are given by $\hat{\imath} = e_1 = (1, 0)$ and $\hat{\jmath} = e_2 = (0, 1)$. Their dot product is $\hat{\imath} \cdot \hat{\jmath} = 0$.

We can generalize this to any angle θ by decomposing $u \in \mathbf{R}^2$ into its standard basis components using the following diagram:





[1] [1], [2], [3], and [4] are all available for free on the internet (see links in the References section).

We consider the angle θ between $\hat{\imath}$ and u. From SOH CAH TOA with H (hypotenuse) $= \|u\| = 1$, we see that $\sin\theta =$ O (opposite) and $\cos\theta =$ A (adjacent). This gives us[2]

$$u = \cos\theta \begin{bmatrix}1\\0\end{bmatrix} + \sin\theta \begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}\cos\theta\\ \sin\theta\end{bmatrix}$$

We see that $\hat{\imath} \cdot u = (1,0)^T(\cos\theta, \sin\theta) = \cos\theta$, and this relationship actually holds for any two unit vectors. Furthermore, from the plot on the right, $\cos\theta > 0$ when $|\theta| < 90^\circ$ and $\cos\theta < 0$ when $|\theta| > 90^\circ$.

The **cosine formula** further generalizes this relationship to arbitrary vectors v and w as follows

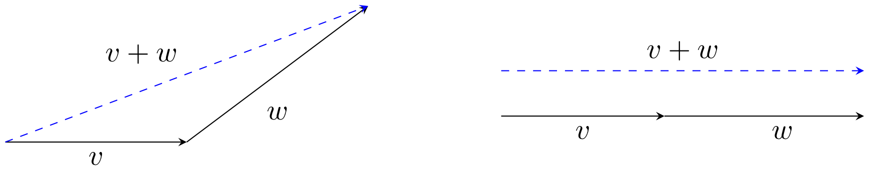
$$\frac{v \cdot w}{\|v\|\|w\|} = \cos\theta, \qquad v \cdot w = \|v\|\|w\|\cos\theta$$

Thus, we have the following relationships for the dot product between any vectors v and w:

$v \cdot w > 0$	:	Angle between vectors is acute ($|\theta| < 90^\circ$)
$v \cdot w = 0$	:	Vectors are orthogonal ($\theta = 90^\circ$)
$v \cdot w < 0$	:	Angle between vectors is obtuse ($|\theta| > 90^\circ$)

Furthermore, since $|\cos\theta| \leq 1$, we see that $|v \cdot w| \leq \|v\|\|w\|$ (**Cauchy-Schwartz**).

From the diagram below, we also see that $\|v + w\| \leq \|v\| + \|w\|$ (**Triangle Inequality**).



Finally, since $v^Tw = w^Tv = 0$ for orthogonal vectors v and w, we have the following:

$$\|v + w\|^2 = (v + w)^T(v + w) = v^Tv + v^Tw + w^Tv + w^Tw = v^Tv + w^Tw = \|v\|^2 + \|w\|^2$$

which is just the **Pythagorean Theorem**. Similarly, we can also show $\|v - w\|^2 = \|v\|^2 + \|w\|^2$.

1.1 Lines and Planes

The equation for a line can be written as $x_2 = ax_1 + d$. Rearranging, we get $x_2 - ax_1 = d$. We can multiply through by an arbitrary constant and define new constants to get $c_1x_1 + c_2x_2 = b$. Therefore, $c^Tx = b$ is the equation for a line when $c, x \in \mathbf{R}^2$. Similarly, $c^Tx = b$ is the equation for a plane when $c, x \in \mathbf{R}^3$.

[2] Note the trigonometric identity $\|u\| = \sin^2\theta + \cos^2\theta = 1$.

2 Matrices

We can see below that multiplying a matrix $A \in \mathbf{R}^{m\times n}$ by a vector $x \in \mathbf{R}^n$ gives a linear combination of the columns of A:

$$Ax = x_1 \begin{bmatrix} | \\ a_1 \\ | \end{bmatrix} + \dots x_n \begin{bmatrix} | \\ a_n \\ | \end{bmatrix}$$

Below are some definitions and rules that are often needed when manipulating matrices:

- $AB \neq BA$ (matrix multiplication does not commute)
- Transpose: $(AB)^T = B^T A^T$
- Exponents: $A^p = AA\dots A, \quad (A^p)^q = (A^q)^p, \quad A^{p+q} = A^p A^q, \quad A^0 = I$

2.1 Pivots and Rank

When performing Gauss-Jordan elimination, the columns that contain the first non-zero entry for a given row are known as **pivot columns**. Any column that is not a pivot column is a **free column** (with an associated **free variable**) and can be written as a linear combination of the pivot columns. Therefore, the number of independent columns in a matrix is the number of pivot columns.

The **rank** r of a matrix A, rank$(A) = r$, is the number of independent columns (number of pivots[3]). For any matrix A, the number of independent columns equals the number of independent rows, or equivalently, rank$(A) =$ rank(A^T).

We say a matrix $A \in \mathbf{R}^{m\times n}$ is **full column rank** if all columns are independent ($r = n$) and **full row rank** if all rows are independent ($r = m$). A square matrix is **full rank** when all columns (and equivalently rows), are independent.

2.2 Inverses

We have the following for the **inverse** of a matrix (note that $A \in \mathbf{R}^{n\times n}$ must be square):

- Calculating the inverse of A: $[A|I_n] \;\rightarrow\; [I_n|A^{-1}]$ (Gauss-Jordan elimination)
- A must have n independent columns to be invertible (all columns have pivots): rank$(A) = n$
- $Ax = b \;\rightarrow\; x = A^{-1}b$ (the matrix is invertible if the system has a solution)
- If A is invertible, then $Ax = \mathbf{0}$ has only one solution (x being the zero vector).
- If A is invertible, then $\det(A) \neq 0$.
- $A^{-1}A = I = AA^{-1}$
- $(AB)^{-1} = B^{-1}A^{-1}$
- $(A^{-1})^T = (A^T)^{-1}$ (often writen A^{-T})
- For diagonal matrix D, the inverse D^{-1} is also diagonal with $1/d_{ii}$ in the diagonal entries.
- A triangular matrix is invertible iff there are no zeros on the diagonal.
- The matrix $A^T A$ is invertible iff A has independent columns.

A square matrix that is not invertible is **singular**. A singular matrix A has more than one solution to $Ax = \mathbf{0}$ and $\det(A) = 0$.

[3] To calculate the rank of a matrix, perform Gauss-Jordan elimination and count the number of pivots.

2.3 Determinants and Trace

The **determinant**, denoted $\det(A)$ or $|A|$, maps a square matrix $A \in \mathbf{R}^{n\times n}$ into a real number. Think of the determinant as a measure of volume: the absolute value of the determinant, $|\det(A)|$, equals the volume of the "box" (parallelepiped) whose edges are formed by the vectors comprising the rows (or columns) of A. Some useful properties:

- $\det(I) = 1$
- Multiplying one row or column of A scales $\det(A)$ by the same amount (similar to volume)
- $\det(\alpha A) = \alpha^n \det(A)$
- $\det(AB) = \det(A)\det(B)$
- $\det(A) = \det(A^T)$
- $\det(A^{-1}) = \frac{1}{\det(A)}$ (since $\det(A^{-1}A) = \det(A^{-1})\det(A) = \det(I) = 1$)
- $\det(A) \neq 0$ iff rank$(A) = n$ (i.e. if A is invertible)
- $\det(A) = 0$ iff A is singular (not invertible)

The **trace** of a square matrix $A \in \mathbf{R}^{n\times n}$ is defined as tr$(A) = a_{11} + \cdots + a_{nn}$ (sum of the diagonal elements of A). Some useful properties:

- tr$(I) = n$
- tr$(A + B) =$ tr$(A) +$ tr(B)
- tr$(A) =$ tr(A^T)
- tr$(ABC) =$ tr$(BCA) =$ tr(CAB)

2.4 Special Matrices

A **symmetric** matrix $S \in \mathbf{R}^{n\times n}$ is any square matrix where $S^T = S$. The inverse of a symmetric matrix is also symmetric since $S^{-1} = (S^T)^{-1} = (S^{-1})^T$. Note that for *any* matrix $A \in \mathbf{R}^{m\times n}$, the matrix $A^TA \in \mathbf{R}^{m\times m}$ is symmetric because $(A^TA)^T = A^T(A^T)^T = A^TA$. Similarly, $AA^T \in \mathbf{R}^{n\times n}$ is also symmetric.

A square matrix $Q \in \mathbf{R}^{n\times n}$ is an **orthogonal matrix** if it has orthonormal columns. Since $q_i^T q_j = 1$ when $i = j$ and 0 otherwise, we have $Q^TQ = I = QQ^T$. Therefore, $Q^T = Q^{-1}$ always holds for orthogonal matrices and the solution to $Qx = b$ is simply $x = Q^Tb$ (we do not need to invert the matrix to solve the system of equations!).

A **rank-one matrix** is any matrix $A \in \mathbf{R}^{m\times n}$ that can be written as $A = uv^T$ (the **outer product** of two vectors), where $u \in \mathbf{R}^m$ and $v \in \mathbf{R}^n$.

For any square matrix A, the **matrix exponential** is defined as

$$e^{At} = I + At + \frac{1}{2}(At)^2 + \frac{1}{6}(At)^3 + \cdots = \sum_{k=0}^{\infty} \frac{1}{k!}(At)^k$$

3 Vector Spaces and Subspaces

A **vector space** is a set of vectors[4] that is **closed** under scalar multiplication and vector addition.[5] By closed, we mean that all linear combinations of vectors in a space stay in (are contained in) the space. Note that *every* space contains the origin, i.e. zero vector (take the scalar multiplier $c = 0$). Vector spaces can also contain **subspace**s, which are themselves spaces. Some examples:

- The space $\mathbf{R}^n$ consists of all column vectors with n components (e.g. $\mathbf{R}^2$ is the xy plane).
- The zero vector $\mathbf{0}$ is itself a subspace.
- Lines through the origin are subspaces.
- Planes through the origin are subspaces.
- The quarter-plane (e.g. quadrant I) is not a subspace ($-1v$ pushes v out of the space).
- A plane in $\mathbf{R}^3$ through the origin is a subspace of $\mathbf{R}^3$.
- A plane in $\mathbf{R}^3$ is not in $\mathbf{R}^2$ (vectors with 3 components cannot be in $\mathbf{R}^2$).
- $\mathbf{R}^2$ is not a subspace of $\mathbf{R}^3$ (vectors with 2 components cannot be in $\mathbf{R}^3$).
- The set of all $x \in \mathbf{R}^n$ whose entries sum to zero (vectors with mean zero) is a subspace.[6]

The **column space** (range), $C(A)$, of a matrix $A \in \mathbf{R}^{m\times n}$ consists of all linear combinations of the columns[7] of A. Taking the set of all $x \in \mathbf{R}^n$, we say that Ax fills the column space of A. We can equivalently say $C(A)$ is the set of all b's that satisfy $Ax = b$. Note that a system $Ax = b$ is only solvable if $b \in C(A)$. Some examples:

- Any column of A is in the column space of A (take $x = e_i$ for any $i = 1, \ldots, n$).
- An $n \times n$ matrix has $C(A) = \mathbf{R}^n$ only when A is invertible.
- $C(A)$ is a subspace, since if $b_1, b_2 \in C(A)$, then there exist vectors x_1 and x_2 such that $Ax_1 = b_1$ and $Ax_2 = b_2$, and therefore, $b_1 + b_2 = Ax_1 + Ax_2 = A(x_1 + x_2) \in C(A)$. Furthermore, $cb_1 = c(Ax_1) = A(cx_1) \in C(A)$. (linear combinations of vectors in the column space stay in the column space).
- The solution, x, to $Ax = b$ is not a subspace because $x = \mathbf{0}$ is not a solution (unless $b = \mathbf{0}$).

The **row space** of a matrix A is the column space of its transpose, $C(A^T)$.

The **nullspace**, $N(A)$, of a matrix $A \in \mathbf{R}^{m\times n}$ consists of all solutions[8] to $Ax = \mathbf{0}$. Some examples:

- $x = \mathbf{0}$ is always in the nullspace of any matrix A.
- For an invertible matrix, $x = \mathbf{0}$ is the *only* solution ($N(A) = \mathbf{0}$ for A invertible).
- $N(A)$ is a subspace, since if $Ax_1 = \mathbf{0}$ and $Ax_2 = \mathbf{0}$, then $A(x_1 + x_2) = Ax_1 + Ax_2 = \mathbf{0} + \mathbf{0} = \mathbf{0}$ and $c(Ax) = c\mathbf{0} = \mathbf{0}$ (linear combinations of vectors in the nullspace stay in the nullspace).

[4] "Vector" can also refer to matrices or even functions (e.g. the space containing all polynomials of degree n).

[5] Spaces must also obey other "common sense" rules (e.g. $x + y = y + x$, $x + 0 = x$, $x + -x = 0$, etc).

[6] It is the nullspace of $A = (1, 1, \ldots, 1)^T$, since all x in the subspace satisfy $Ax = \mathbf{0}$.

[7] Note that $C(A) \in \mathbf{R}^m$ (not $\mathbf{R}^n$!) because Ax gives a vector in $\mathbf{R}^m$, where m = number of rows.

[8] Note that $N(A) \in \mathbf{R}^n$ since $x \in \mathbf{R}^n$ for Ax, where n = number of columns.

3.1 Independence

The sequence of vectors $v_1, \ldots, v_n$ is **linearly independent** if $c_1 v_1 + \cdots + c_n v_n = \mathbf{0}$ only when $c_0, \ldots, c_n$ are all zeros. Intuitively, vectors $v_1, \ldots, v_n$ are independent when no vector can be written as a combination of the others. Conversely, vectors $v_1, \ldots, v_n$ are **linearly dependent** when at least one vector is a combination of the others.

For matrices, the columns of A are linearly independent when $N(A)$ contains only the zero vector (i.e. the only solution to $Ax = 0$ is $x = \mathbf{0}$). This is equivalent to the matrix having full column rank[9]: $r = n$. Similarly, the columns of A are dependent when there is a non-zero vector in the nullspace (i.e. rank $< n$).

Some examples:

- If three vectors in $\mathbf{R}^3$ are not in the same plane, they are linearly independent.
- If three vectors in $\mathbf{R}^3$ are in the same plane, they are linearly dependent.
- The vectors $(1, 0)$ and $(0, 1)$ are independent
- The vectors $(1, 1)$ and $(-1, -1)$ are dependent (they lie on a line through the origin).
- The vectors $(1, 1)$ and $(0, 0)$ are dependent (zero vector is always linearly dependent).
- Any sequence of n vectors in $\mathbf{R}^m$ will be linearly dependent for $m < n$ (i.e. the columns of any matrix with more rows than columns will be dependent).

3.2 Span

If we start with any set $\mathbf{S}$ of vectors in a vector space $\mathbf{V}$ (e.g. take two arbitrary vectors in $\mathbf{V} = \mathbf{R}^3$) and take all linear combinations of these vectors, we get a subspace $\mathbf{V}_s$ of $\mathbf{V}$. We say that $\mathbf{V}_s$ is **spanned** by $\mathbf{S}$ (or $\mathbf{V}_s$ is the **span** of $\mathbf{S}$), which means that any vector in $\mathbf{V}_s$ can be represented by a linear combination of the vectors in $\mathbf{S}$. In other words, a set of vectors **spans** a space if all linear combinations of the vectors fill the space.

Note that the columns of a matrix span its column space. They might be dependent.

3.3 Orthogonality

Two subspaces V and W of a vector space are **orthogonal subspaces** if any vector $v \in V$ and any vector $w \in W$ satisfy $v^T w = 0$. We also say that W is the **orthogonal complement** of V. Some examples:

- The zero vector is the only vector that is in all subspaces and is orthogonal to every subspace.
- If the floor of a room (extended to infinity) represents a subspace and the wall of a room another subspace, they are not orthogonal because any non-vertical line drawn on the wall is not orthogonal to the line where the wall and floor meet. Furthermore, the line where the wall and floor meet is also in both subspaces, so those subspaces cannot be orthogonal.
- The floor of a room is an orthogonal subspace to the line formed where two walls meet. In this case, the line goes through the plane at 90°.

[9] To check for independence, perform Gauss-Jordan elimination and confirm that every column has a pivot.

- The row space and nullspace of a matrix are orthogonal subspaces (orthogonal complements) in $\mathbf{R}^n$

The see the last point, note that every vector $x \in \mathbf{R}^n$ in the nullspace of A is orthogonal to every row in A, since $(-\text{row}_i-)^T x = 0$ is always satisfied for $Ax = \mathbf{0}$. Therefore, any linear combination of the dot products of each row with x is also 0, and we have that the row space and nullspace of any matrix are orthogonal. This can also be seen by noting that the row space satisfies $A^T y \subseteq \mathbf{R}^n$ for any $y \in \mathbf{R}^m$ and the nullspace satisfies $Ax = \mathbf{0}$ for any $x \in \mathbf{R}^n$. Therefore, $(A^T y)^T x = y^T(Ax) = y^T\mathbf{0} = 0$ and we see that the subspaces are orthogonal.

3.4 Basis

A **basis** for a vector space is a sequence of vectors that are linearly independent and span the space. Since the basis vectors are independent, there is only one way to write a vector as a combination of the basis vectors. For a given space, the basis is not unique, but every basis has the same number of vectors. Some examples:

- The $n \times n$ identity matrix I is the **standard basis** for $\mathbf{R}^n$.
- The pivot columns of any A are a basis for its column space.[10]
- The columns of *every* invertible matrix ($n \times n$) give a basis for $\mathbf{R}^n$. Thus, there are infinitely many bases for $\mathbf{R}^n$.
- For any matrix A with independent columns, we can construct[11] a new matrix B containing orthonormal columns (i.e. orthonormal basis vectors) such that $C(A) = C(B)$.
- Two independent vectors span a plane and are a basis for that plane. If we add one more vector in the plane, all three vectors will still span the plane, but they are not a basis because the three vectors are not independent.

3.5 Dimension

The **dimension** of a space is the number of basis vectors in every basis. We may think of the dimension as the "degrees of freedom" in the space.

The dimension of the column space (and row space) of a matrix A is equal to $r = \text{rank}(A)$ (number of independent columns or pivots or rows). The dimension of the nullspace is equal to $n - r$ (number of free columns).

$$\boxed{\begin{aligned} r &= \text{rank}(A) && \text{(Number of pivot columns)} \\ &= \dim C(A) && \text{(Number of basis vectors in the column space)} \\ &= \dim C(A^T) && \text{(Number of basis vectors in the row space)} \end{aligned}}$$

Note that if you know the dimension of a space and you have r independent vectors that are in the space, then those vectors form a basis for that space.

[10] Perform Gauss-Jordan elimination and take the original columns of A corresponding to the pivot columns.

[11] For instance, using the **Gram-Schmidt** process, which is a generalization of projections onto lines.

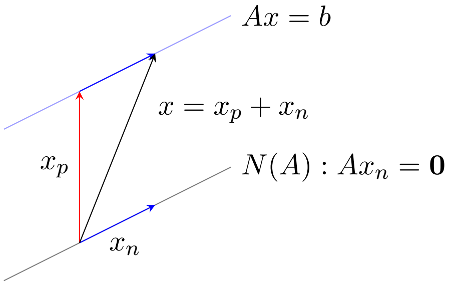
4 Solutions of Linear Systems

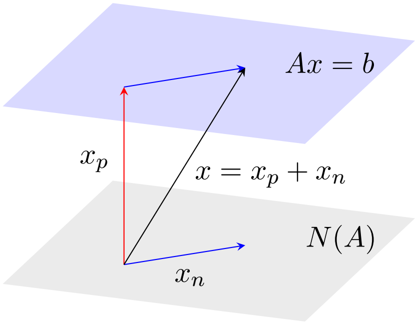
A solution, x, of any linear system $Ax = b$ can always be expressed as the sum of:

- A particular solution, x_p, satisfying $Ax_p = b$. (x_p is not necessarily unique)
- Any vector in the nullspace, $x_n \in N(A)$, satisfying $Ax_n = \mathbf{0}$.

This is the case because $A(x_p + x_n) = Ax_p + Ax_n = b + \mathbf{0} = b$. Note that we are free to attach a scalar to x_n, since cx_n stays in the nullspace, but we cannot attach a scalar to the particular solution because $A(cx_p) = c(Ax_p) = cb \neq b$.

The complete solution is visualized in the figures below. If the nullspace of A is 1-dimensional (i.e. there is only one free column in A), the nullspace takes the form of a line, as shown on the left. In this case, the particular solution shifts the nullspace away from the origin and any point on the shifted line satisfies $Ax = b$.





When the nullspace has more than one dimension, it can be expressed as a linear combination of $n - r$ independent vectors, which fills out a (hyper) plane through the origin, as shown on the right for a two dimensional nullspace. In this case, we see that the particular solution shifts the nullspace up to the plane satisfying $Ax = b$. Any value of x in this plane is a solution to $Ax = b$.

A special case of the complete solution is when A is invertible (square, full rank). In this case, $x_n = \mathbf{0}$ (the nullspace contains only the zero vector) and the solution to $Ax = b$ is just the (unique) particular solution. Furthermore, a solution always exists, since $C(A) = \mathbf{R}^n$ (all n-dimensional vectors are in the column space, so any b can be represented as a linear combination of the columns of A).

More generally, for a matrix $A \in \mathbf{R}^{m \times n}$, the table below shows all possibilities for solutions:

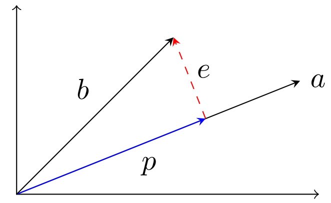
Rank	Description	$Ax = b$
$r = m$ and $r = n$	Full rank (square and invertible)	1 solution
$r = m$ and $r < n$	Full row rank (short and wide, $m < n$, underdetermined)	∞ solutions
$r < m$ and $r = n$	Full column rank (tall and thin, $m > n$, overdetermined)	0 or 1 solution
$r < m$ and $r < n$	Not full rank	0 or ∞ solutions

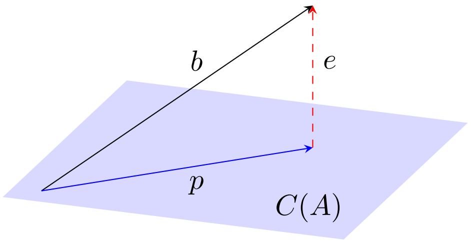
5 Projections and Least Squares

The projection of a vector b onto a line spanned by a vector a is found using orthogonal vectors. From the diagram on the left, we see that error vector $e = b - p$ is orthogonal to a. Therefore, $a^T(b-p) = 0$, where $p = \hat{x}a$ is some multiple of a. Solving for $\hat{x}$, and subsequently p, we get

$$a^T(b - \hat{x}a) = 0 \quad \rightarrow \quad \hat{x} = \frac{a^Tb}{a^Ta}, \qquad p = \hat{x}a = \frac{aa^T}{a^Ta}b.$$

We call $P = (aa^T)/(a^Ta)$ the **projection matrix** (projector). Note that for the case of projecting onto a line, this is a rank-one matrix, since it is a column vector times a row vector (divided by a scalar). Therefore, P projects onto a one-dimensional subspace (i.e. the column space of P is the line through a). Furthermore, since $(I-P)b = b - p$, we see that $I - P$ is also a projection matrix and projects onto the line perpendicular to a.





The projection of a vector b onto the (hyper) plane spanned by the columns of A (i.e. the column space of A) follows a similar approach and is illustrated in the diagram on the right. Thus, we wish to solve for $\hat{x}$ that gives us $A\hat{x} = p$, where p is a vector in $C(A)$ that is closest to b given the the error vector $e = b - p$. Since e is orthogonal to $C(A)$, we must have[12] $A^T(b - A\hat{x}) = \mathbf{0}$ or $A^TA\hat{x} = A^Tb$ (which is just $Ax = b$ with both sides left multiplied by A^T). Note that A^TA is square symmetric and if the columns of A are independent, then A^TA is invertible. This gives us the following solution and corresponding projection matrix, which projects any vector b onto the column space of A:

$$\boxed{A^TA\hat{x} = A^Tb, \qquad \hat{x} = (A^TA)^{-1}A^Tb, \qquad P = A(A^TA)^{-1}A^T}$$

We also note that $P^2 = P$ since $P^2 = A(A^TA)^{-1}(A^TA)(A^TA)^{-1}A^T = A(A^TA)^{-1}A^T = P$. This is a general property of projection matrices. Intuitively, the projection of a vector that is already in the column space of A does not change the vector ($p = Pb = P(Pb) = p$).

If the columns of matrix A are orthogonal, we have $A^TA = I$, leading to $\hat{x} = A^Tb$ and $P = AA^T$. Furthermore, if A is an orthogonal matrix (square with orthonormal columns), then $\hat{x} = Q^Tb$ and $P = I$. Intuitively, since the column space of Q is $C(Q) = \mathbf{R}^m$, every b is in $C(Q)$, and therefore the projection leaves b unchanged.

[12] Note that since $A^Te = \mathbf{0}$, we equivalently have that vector e is in the nullspace of A^T.

5.1 Least Squares

The above solutions can also be derived using calculus. To start, we have the following rules for gradients (the vector of partial derivatives with respect to each vector component):

$$\boxed{\frac{\partial}{\partial x}x^Tx = 2x, \qquad \frac{\partial}{\partial x}Ax = A^T, \qquad \frac{\partial}{\partial x}x^TAx = (A + A^T)x}$$

Note that for symmetric matrices, $S = S^T$, we see that the partial derivatives of Sx and x^TSx are simply S and $2Sx$, respectively.

We wish to minimize the difference between b and some vector in the column space of A such that $||A\hat{x} - b||^2$ is minimized. Note that

$$\begin{aligned} ||A\hat{x} - b||^2 &= (A\hat{x} - b)^T(A\hat{x} - b) \\ &= \hat{x}^TA^TA\hat{x} - \hat{x}^TA^Tb - b^TA\hat{x} + b^Tb \\ &= \hat{x}^TA^TA\hat{x} - 2\hat{x}^TA^Tb + b^Tb \end{aligned}$$

where we used the fact that $\hat{x}^TA^Tb \in \mathbf{R}$ is a scalar and therefore its transpose, $b^TA\hat{x}$, is equal to itself. Taking the gradient and setting it to zero, we get

$$\frac{\partial}{\partial x}\left(\hat{x}^TA^TA\hat{x} - 2\hat{x}^TA^Tb + b^Tb\right) = 2A^TA\hat{x} - 2A^Tb = 0 \quad \rightarrow \quad A^TA\hat{x} = A^Tb$$

which is the same as the result derived using projection matrices.

6 Eigenvectors and Eigenvalues

Almost all vectors change direction when multiplied by a matrix, but some do not. Vectors x that point in the same direction as Ax are **eigenvector**s. Eigenvectors x lie along the same line as Ax. We express this as $Ax = \lambda x$, where A is an $n \times n$ square matrix, x is an eigenvector, and λ is the **eigenvalue** associated with x. The eigenvalue measures how much the eigenvector is stretched/shrunk when multiplied by A:

- $|\lambda| > 1$: x is stretched by A
- $|\lambda| < 1$: x is shrunk by A
- $\lambda < 0$: The direction of x is reversed (but it still remains on the same line as x).
- $\lambda = 0$: x is in the nullspace[13] of A

Since $Ax = \lambda x$ can be rewritten as $(A - \lambda I)x = 0$, we see that eigenvectors make up the nullspace of the matrix $(A - \lambda I)$. Furthermore, if there is a nonzero solution, then $(A - \lambda I)$ is not invertible (i.e. it is singular) and therefore $\det(A - \lambda I) = 0$. Writing out the determinant gives us the **characteristic equation** (polynomial of degree n) and solving for the roots gives us the eigenvalues. Note that there are n roots (eigenvalues) for an $n \times n$ matrix.[14] For each eigenvalue, λ_i, the matrix $(A - \lambda_i I)$ is singular (there is a nonzero x_i in the nullspace), so we can solve for x_i that gives

[13] This is still considered to be on the line as x, since every vector passes through the origin.

[14] But multiple eigenvalues can have the same value. Eigenvalues can also be complex (conjugates) even if A is real.

$(A - \lambda_i I)x_i = 0$. These x_i are the eigenvectors.

For example, for a 2×2 matrix, we have[15]

$$\det(A - \lambda I) = \begin{vmatrix} a-\lambda & b \\ c & d-\lambda \end{vmatrix} = (a-\lambda)(d-\lambda) - bc = \lambda^2 - ad\lambda + (ad - bc) = 0$$

Solving for the roots of this characteristic polynomial gives us the two eigenvalues λ_1 and λ_2. Plugging each λ_i into $(A - \lambda_i I)x_i = 0$ and solving each system of equations gives us two eigenvectors x_1 and x_2.

Note that eigenvectors are not unique. If we multiply both sides of $Ax = \lambda x$ by a scalar c, we get $c(Ax) = c(\lambda x) \to A(cx) = \lambda(cx)$, so cx is also a valid eigenvector, albeit one that still lies on the same line as x. Eigenvectors are commonly scaled so that they are all unit vectors.

Some useful properties of eigenvalues and eigenvectors:

- $\text{tr}(A) = \lambda_1 + \cdots + \lambda_n$ (the sum of the eigenvalues equals the trace)
- $\det(A) = \lambda_1 \cdots \lambda_n$ (the product of the eigenvalues equals the determinant)
- The diagonal entries of a triangular matrix are its eigenvalues.
- $(cA)x = (c\lambda)x$ (scaling a matrix leaves the eigenvectors alone but scales the eigenvalues)
- $A^k x = \lambda^k x$ (e.g. for $k = 2$, we have $A^2x = AAx = A(\lambda x) = \lambda^2 x$)
- If A is singular (not invertible), at least one $\lambda = 0$ (other λ's can be nonzero)
- If A is invertible, $\lambda = 0$ cannot be an eigenvalue (we shift A by λ to make it singular)
- $A^{-1}x = \lambda^{-1}x$ (the eigenvalues of the inverse of a matrix are the inverses of the eigenvalues)

Most matrices $A \in \mathbf{R}^{n \times n}$ have n eigenvector directions and n distinct eigenvalues, but some types of matrices have unique properties:

- Symmetric matrices have orthogonal eigenvectors (which can be scaled to be orthonormal)
- Identity matrices ($Ix = x$): All vectors are eigenvectors of I and all λ's are 1
- Projection matrices:
  - If x is in the column space, then $Px = x$ and $\lambda = 1$ (the column space projects onto itself, so projection does not change x).
  - If x is orthogonal to the column space, then $Px = 0$ and $\lambda = 0$ (x is in the null space).
  - Projection matrices (and only projection matrices) only have eigenvalues that are either 0 or 1.

6.1 Diagonalizing a Matrix

We may write $Ax_i = \lambda_i x_i$ in matrix form by placing the all eigenvectors into the columns of a matrix X and all the the eigenvalues into a diagonal marix Λ as follows

$$AX = A \begin{bmatrix} | & & | \\ x_1 & \cdots & x_n \\ | & & | \end{bmatrix} = \begin{bmatrix} | & & | \\ x_1 & \cdots & x_n \\ | & & | \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} = X\Lambda$$

[15] Note that this is equal to $\lambda^2 - \text{tr}(A)\lambda + \det(A)$

where we refer to X as the eigenvector matrix and Λ as the eigenvalue matrix.

If all the eigenvalues in Λ are different, then all eigenvectors are guaranteed to be independent. If there are repeated eigenvalues, then we must check the eigenvectors to see if they are independent.[16]

If a real matrix $A \in \mathbf{R}^{n\times n}$ has linearly independent eigenvectors $x_1, \cdots, x_n$, then the eigenvector matrix X is invertible and we say that A is **diagonalizable**.[17] Diagonalizable matrices can therefore be written as $A = X\Lambda X^{-1}$ or alternatively $\Lambda = X^{-1}AX$.

Note that since $A^2 = AA = (X\Lambda X^{-1})(X\Lambda X^{-1}) = X\Lambda(X^{-1}X)\Lambda X^{-1} = X\Lambda^2 X^{-1}$, squaring a matrix squares the eigenvalues but leaves the eigenvectors unchanged. More generally, $A^k = X\Lambda^k X^{-1}$.

Similarly, for any diagonalizable matrix A, we have

$$e^{At} = \sum_{k=0}^{\infty} \frac{1}{k!}(At)^k = \sum_{k=0}^{\infty} \frac{1}{k!} X(\Lambda t)^k X^{-1} = X\left(\sum_{k=0}^{\infty} \frac{1}{k!}(\Lambda t)^k\right) X^{-1} = Xe^{\Lambda t}X^{-1}$$

6.2 Symmetric and Positive Definite Matrices

Symmetric matrices have only *real* eigenvalues and *orthogonal* eigenvectors, which can be normalized to be *orthonormal*. The **spectral theorem** (principal axis theorem) says that every symmetric matrix has factorization $S = Q\Lambda Q^T$, where the eigenvector matrix ($X = Q$) is orthogonal ($Q^{-1} = Q^T$).[18] Expanding out the factorization, we have[19]:

$$\begin{aligned} S &= Q\Lambda Q^T = \begin{bmatrix} | & & | \\ x_1 & \cdots & x_n \\ | & & | \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \begin{bmatrix} — x_1 — \\ \vdots \\ — x_n — \end{bmatrix} = \begin{bmatrix} | & & | \\ \lambda_1 x_1 & \cdots & \lambda_n x_n \\ | & & | \end{bmatrix} \begin{bmatrix} — x_1 — \\ \vdots \\ — x_n — \end{bmatrix} \\ &= \lambda_1 x_1 x_1^T + \cdots + \lambda_n x_n x_n^T = \lambda_1 P_1 + \cdots + \lambda_n P_n \end{aligned}$$

where $P_i = x_i x_i^T$ are rank-one projection matrices, and we see that a symmetric matrix A can be written as a linear combination of projection matrices weighted by their eigenvalues.

A symmetric matrix with all *positive* real eigenvalues is a **positive definite** matrix. Equivalently, a matrix A is positive definite if $x^T Ax > 0$ for every non-zero vector x.[20] If the eigenvalues are non-negative (i.e. $\lambda_i \geq 0$), the matrix is **positive semidefinite** and $x^T Ax \geq 0$.

If A and B are positive (semi) definite, so is $A + B$ because $x^T(A + B)x = x^T Ax + x^T Bx > 0$.

[16] For instance, the identity matrix has all $\lambda_i = 1$, but the eigenvectors are independent (X can be *any* basis in $\mathbf{R}^n$ for the identity matrix).

[17] If the eigenvectors are not all independent, then matrix A is not diagonalizable.

[18] Note that $Q\Lambda Q^T$ is symmetric since $(Q\Lambda Q^T)^T = (Q^T)^T \Lambda^T Q = Q\Lambda Q^T$.

[19] Try this with 2×2 matrices for Q and Λ to convince yourself it holds.

[20] A matrix is also positive definite if all principal minors (upper left determinants) are positive: $a_{11} > 0, a_{11}a_{22} - a_{21}a_{12} > 0, \cdots, \det(A) > 0$

For any matrix R with independent columns, the symmetric matrix $A = R^T R$ is positive definite since $x^T Ax = x^T R^T Rx = (Rx)^T (Rx) = \|Rx\|^2 > 0$ (because square of real number is positive).

It can also be shown that every positive definite matrix A can be written as $A = LL^T$, where L is lower triangular (**Cholesky decomposition**).

References

[1] Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*, Cambridge University Press.

[2] Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.*, 2nd Ed., Springer.

[3] Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*, MIT Press.

[4] Strang, G. (2005). MIT 18.06 Linear Algebra [Video lectures]. YouTube. https://www.youtube.com/playlist?list=PLE7DDD91010BC51F8

[5] Strang, G. (2016). *Introduction to Linear Algebra*, 5th Ed., Wellesley-Cambridge Press.