Linear Algebra Review and Reference

Zico Kolter (updated by Chuong Do)

September 30, 2015

Contents

1 Basic Concepts and Notation

Linear algebra provides a way of compactly representing and operating on sets of linear equations. For example, consider the following system of equations:

$$\begin{array}{rcrcl} 4x_1 & - & 5x_2 & = & -13 \\ -2x_1 & + & 3x_2 & = & 9. \end{array}$$

This is two equations and two variables, so as you know from high school algebra, you can find a unique solution for x_1 and x_2 (unless the equations are somehow degenerate, for example if the second equation is simply a multiple of the first, but in the case above there is in fact a unique solution). In matrix notation, we can write the system more compactly as

$$Ax = b$$

with

$$A = \begin{bmatrix} 4 & -5 \\ -2 & 3 \end{bmatrix}, \quad b = \begin{bmatrix} -13 \\ 9 \end{bmatrix}.$$

As we will see shortly, there are many advantages (including the obvious space savings) to analyzing linear equations in this form.

1.1 Basic Notation

We use the following notation:

- By $A \in \mathbb{R}^{m \times n}$ we denote a matrix with m rows and n columns, where the entries of A are real numbers.
- By $x \in \mathbb{R}^n$, we denote a vector with n entries. By convention, an n-dimensional vector is often thought of as a matrix with n rows and 1 column, known as a ***column vector***. If we want to explicitly represent a ***row vector*** — a matrix with 1 row and n columns — we typically write x^T (here x^T denotes the transpose of x, which we will define shortly).
- The ith element of a vector x is denoted x_i:

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

- We use the notation a_{ij} (or A_{ij}, $A_{i,j}$, etc) to denote the entry of A in the ith row and jth column:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

- We denote the jth column of A by a_j or $A_{:,j}$:

$$A = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix}.$$

- We denote the ith row of A by a_i^T or $A_{i,:}$:

$$A = \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix}.$$

- Note that these definitions are ambiguous (for example, the a_1 and a_1^T in the previous two definitions are *not* the same vector). Usually the meaning of the notation should be obvious from its use.

2 Matrix Multiplication

The product of two matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$ is the matrix

$$C = AB \in \mathbb{R}^{m \times p},$$

where

$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}.$$

Note that in order for the matrix product to exist, the number of columns in A must equal the number of rows in B. There are many ways of looking at matrix multiplication, and we'll start by examining a few special cases.

2.1 Vector-Vector Products

Given two vectors $x, y \in \mathbb{R}^n$, the quantity $x^T y$, sometimes called the ***inner product*** or ***dot product*** of the vectors, is a real number given by

$$x^T y \in \mathbb{R} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} x_i y_i.$$

Observe that inner products are really just special case of matrix multiplication. Note that it is always the case that $x^T y = y^T x$.

Given vectors $x \in \mathbb{R}^m$, $y \in \mathbb{R}^n$ (not necessarily of the same size), $xy^T \in \mathbb{R}^{m \times n}$ is called the ***outer product*** of the vectors. It is a matrix whose entries are given by $(xy^T)_{ij} = x_i y_j$, i.e.,

$$xy^T \in \mathbb{R}^{m \times n} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_m y_1 & x_m y_2 & \cdots & x_m y_n \end{bmatrix}.$$

As an example of how the outer product can be useful, let $\mathbf{1} \in \mathbb{R}^n$ denote an n-dimensional vector whose entries are all equal to 1. Furthermore, consider the matrix $A \in \mathbb{R}^{m \times n}$ whose columns are all equal to some vector $x \in \mathbb{R}^m$. Using outer products, we can represent A compactly as,

$$A = \begin{bmatrix} | & | & & | \\ x & x & \cdots & x \\ | & | & & | \end{bmatrix} = \begin{bmatrix} x_1 & x_1 & \cdots & x_1 \\ x_2 & x_2 & \cdots & x_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_m & x_m & \cdots & x_m \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} = x\mathbf{1}^T.$$

2.2 Matrix-Vector Products

Given a matrix $A \in \mathbb{R}^{m \times n}$ and a vector $x \in \mathbb{R}^n$, their product is a vector $y = Ax \in \mathbb{R}^m$. There are a couple ways of looking at matrix-vector multiplication, and we will look at each of them in turn.

If we write A by rows, then we can express Ax as,

$$y = Ax = \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix} x = \begin{bmatrix} a_1^T x \\ a_2^T x \\ \vdots \\ a_m^T x \end{bmatrix}.$$

In other words, the ith entry of y is equal to the inner product of the ith *row* of A and x, $y_i = a_i^T x$.

Alternatively, let's write A in column form. In this case we see that,

$$y = Ax = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \\ a_1 \\ \\ \end{bmatrix} x_1 + \begin{bmatrix} \\ a_2 \\ \\ \end{bmatrix} x_2 + \ldots + \begin{bmatrix} \\ a_n \\ \\ \end{bmatrix} x_n \ .$$

In other words, y is a ***linear combination*** of the *columns* of A, where the coefficients of the linear combination are given by the entries of x.

So far we have been multiplying on the right by a column vector, but it is also possible to multiply on the left by a row vector. This is written, $y^T = x^T A$ for $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^m$, and $y \in \mathbb{R}^n$. As before, we can express y^T in two obvious ways, depending on whether we express A in terms on its rows or columns. In the first case we express A in terms of its columns, which gives

$$y^T = x^T A = x^T \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} x^T a_1 & x^T a_2 & \cdots & x^T a_n \end{bmatrix}$$

which demonstrates that the ith entry of y^T is equal to the inner product of x and the ith *column* of A.

Finally, expressing A in terms of rows we get the final representation of the vector-matrix product,

$$\begin{aligned} y^T &= x^T A \\ &= \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix} \\ &= x_1 \begin{bmatrix} — & a_1^T & — \end{bmatrix} + x_2 \begin{bmatrix} — & a_2^T & — \end{bmatrix} + \ldots + x_n \begin{bmatrix} — & a_n^T & — \end{bmatrix} \end{aligned}$$

so we see that y^T is a linear combination of the *rows* of A, where the coefficients for the linear combination are given by the entries of x.

2.3 Matrix-Matrix Products

Armed with this knowledge, we can now look at four different (but, of course, equivalent) ways of viewing the matrix-matrix multiplication $C = AB$ as defined at the beginning of this section.

First, we can view matrix-matrix multiplication as a set of vector-vector products. The most obvious viewpoint, which follows immediately from the definition, is that the (i, j)th

entry of C is equal to the inner product of the ith row of A and the jth column of B. Symbolically, this looks like the following,

$$C = AB = \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix} \begin{bmatrix} | & | & & | \\ b_1 & b_2 & \cdots & b_p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} a_1^T b_1 & a_1^T b_2 & \cdots & a_1^T b_p \\ a_2^T b_1 & a_2^T b_2 & \cdots & a_2^T b_p \\ \vdots & \vdots & \ddots & \vdots \\ a_m^T b_1 & a_m^T b_2 & \cdots & a_m^T b_p \end{bmatrix}.$$

Remember that since $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$, $a_i \in \mathbb{R}^n$ and $b_j \in \mathbb{R}^n$, so these inner products all make sense. This is the most "natural" representation when we represent A by rows and B by columns. Alternatively, we can represent A by columns, and B by rows. This representation leads to a much trickier interpretation of AB as a sum of outer products. Symbolically,

$$C = AB = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} — & b_1^T & — \\ — & b_2^T & — \\ & \vdots & \\ — & b_n^T & — \end{bmatrix} = \sum_{i=1}^{n} a_i b_i^T \ .$$

Put another way, AB is equal to the sum, over all i, of the outer product of the ith column of A and the ith row of B. Since, in this case, $a_i \in \mathbb{R}^m$ and $b_i \in \mathbb{R}^p$, the dimension of the outer product $a_i b_i^T$ is $m \times p$, which coincides with the dimension of C. Chances are, the last equality above may appear confusing to you. If so, take the time to check it for yourself!

Second, we can also view matrix-matrix multiplication as a set of matrix-vector products. Specifically, if we represent B by columns, we can view the columns of C as matrix-vector products between A and the columns of B. Symbolically,

$$C = AB = A \begin{bmatrix} | & | & & | \\ b_1 & b_2 & \cdots & b_p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ Ab_1 & Ab_2 & \cdots & Ab_p \\ | & | & & | \end{bmatrix}.$$

Here the ith column of C is given by the matrix-vector product with the vector on the right, $c_i = Ab_i$. These matrix-vector products can in turn be interpreted using both viewpoints given in the previous subsection. Finally, we have the analogous viewpoint, where we represent A by rows, and view the rows of C as the matrix-vector product between the rows of A and C. Symbolically,

$$C = AB = \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix} B = \begin{bmatrix} — & a_1^T B & — \\ — & a_2^T B & — \\ & \vdots & \\ — & a_m^T B & — \end{bmatrix}.$$

Here the ith row of C is given by the matrix-vector product with the vector on the left, $c_i^T = a_i^T B$.

It may seem like overkill to dissect matrix multiplication to such a large degree, especially when all these viewpoints follow immediately from the initial definition we gave (in about a line of math) at the beginning of this section. However, virtually all of linear algebra deals with matrix multiplications of some kind, and it is worthwhile to spend some time trying to develop an intuitive understanding of the viewpoints presented here.

In addition to this, it is useful to know a few basic properties of matrix multiplication at a higher level:

- Matrix multiplication is associative: $(AB)C = A(BC)$.
- Matrix multiplication is distributive: $A(B + C) = AB + AC$.
- Matrix multiplication is, in general, *not* commutative; that is, it can be the case that $AB \neq BA$. (For example, if $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times q}$, the matrix product BA does not even exist if m and q are not equal!)

If you are not familiar with these properties, take the time to verify them for yourself. For example, to check the associativity of matrix multiplication, suppose that $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times p}$, and $C \in \mathbb{R}^{p\times q}$. Note that $AB \in \mathbb{R}^{m\times p}$, so $(AB)C \in \mathbb{R}^{m\times q}$. Similarly, $BC \in \mathbb{R}^{n\times q}$, so $A(BC) \in \mathbb{R}^{m\times q}$. Thus, the dimensions of the resulting matrices agree. To show that matrix multiplication is associative, it suffices to check that the (i, j)th entry of $(AB)C$ is equal to the (i, j)th entry of $A(BC)$. We can verify this directly using the definition of matrix multiplication:

$$\begin{aligned}
((AB)C)_{ij} &= \sum_{k=1}^{p}(AB)_{ik}C_{kj} = \sum_{k=1}^{p}\left(\sum_{l=1}^{n}A_{il}B_{lk}\right)C_{kj} \\
&= \sum_{k=1}^{p}\left(\sum_{l=1}^{n}A_{il}B_{lk}C_{kj}\right) = \sum_{l=1}^{n}\left(\sum_{k=1}^{p}A_{il}B_{lk}C_{kj}\right) \\
&= \sum_{l=1}^{n}A_{il}\left(\sum_{k=1}^{p}B_{lk}C_{kj}\right) = \sum_{l=1}^{n}A_{il}(BC)_{lj} = (A(BC))_{ij}.
\end{aligned}$$

Here, the first and last two equalities simply use the definition of matrix multiplication, the third and fifth equalities use the distributive property for *scalar multiplication over addition*, and the fourth equality uses the *commutative and associativity of scalar addition*. This technique for proving matrix properties by reduction to simple scalar properties will come up often, so make sure you're familiar with it.

3 Operations and Properties

In this section we present several operations and properties of matrices and vectors. Hopefully a great deal of this will be review for you, so the notes can just serve as a reference for these topics.

3.1 The Identity Matrix and Diagonal Matrices

The ***identity matrix***, denoted $I \in \mathbb{R}^{n \times n}$, is a square matrix with ones on the diagonal and zeros everywhere else. That is,

$$I_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

It has the property that for all $A \in \mathbb{R}^{m \times n}$,

$$AI = A = IA.$$

Note that in some sense, the notation for the identity matrix is ambiguous, since it does not specify the dimension of I. Generally, the dimensions of I are inferred from context so as to make matrix multiplication possible. For example, in the equation above, the I in $AI = A$ is an $n \times n$ matrix, whereas the I in $A = IA$ is an $m \times m$ matrix.

A ***diagonal matrix*** is a matrix where all non-diagonal elements are 0. This is typically denoted $D = \text{diag}(d_1, d_2, \ldots, d_n)$, with

$$D_{ij} = \begin{cases} d_i & i = j \\ 0 & i \neq j \end{cases}$$

Clearly, $I = \text{diag}(1, 1, \ldots, 1)$.

3.2 The Transpose

The ***transpose*** of a matrix results from "flipping" the rows and columns. Given a matrix $A \in \mathbb{R}^{m \times n}$, its transpose, written $A^T \in \mathbb{R}^{n \times m}$, is the $n \times m$ matrix whose entries are given by

$$(A^T)_{ij} = A_{ji}.$$

We have in fact already been using the transpose when describing row vectors, since the transpose of a column vector is naturally a row vector.

The following properties of transposes are easily verified:

- $(A^T)^T = A$
- $(AB)^T = B^T A^T$
- $(A + B)^T = A^T + B^T$

3.3 Symmetric Matrices

A square matrix $A \in \mathbb{R}^{n \times n}$ is ***symmetric*** if $A = A^T$. It is ***anti-symmetric*** if $A = -A^T$. It is easy to show that for any matrix $A \in \mathbb{R}^{n \times n}$, the matrix $A + A^T$ is symmetric and the

matrix $A - A^T$ is anti-symmetric. From this it follows that any square matrix $A \in \mathbb{R}^{n \times n}$ can be represented as a sum of a symmetric matrix and an anti-symmetric matrix, since

$$A = \frac{1}{2}(A + A^T) + \frac{1}{2}(A - A^T)$$

and the first matrix on the right is symmetric, while the second is anti-symmetric. It turns out that symmetric matrices occur a great deal in practice, and they have many nice properties which we will look at shortly. It is common to denote the set of all symmetric matrices of size n as $\mathbb{S}^n$, so that $A \in \mathbb{S}^n$ means that A is a symmetric $n \times n$ matrix;

3.4 The Trace

The ***trace*** of a square matrix $A \in \mathbb{R}^{n \times n}$, denoted $\mathrm{tr}(A)$ (or just $\mathrm{tr}A$ if the parentheses are obviously implied), is the sum of diagonal elements in the matrix:

$$\mathrm{tr}A = \sum_{i=1}^{n} A_{ii}.$$

As described in the CS229 lecture notes, the trace has the following properties (included here for the sake of completeness):

- For $A \in \mathbb{R}^{n \times n}$, $\mathrm{tr}A = \mathrm{tr}A^T$.
- For $A, B \in \mathbb{R}^{n \times n}$, $\mathrm{tr}(A + B) = \mathrm{tr}A + \mathrm{tr}B$.
- For $A \in \mathbb{R}^{n \times n}$, $t \in \mathbb{R}$, $\mathrm{tr}(tA) = t\ \mathrm{tr}A$.
- For A, B such that AB is square, $\mathrm{tr}AB = \mathrm{tr}BA$.
- For A, B, C such that ABC is square, $\mathrm{tr}ABC = \mathrm{tr}BCA = \mathrm{tr}CAB$, and so on for the product of more matrices.

As an example of how these properties can be proven, we'll consider the fourth property given above. Suppose that $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times m}$ (so that $AB \in \mathbb{R}^{m \times m}$ is a square matrix). Observe that $BA \in \mathbb{R}^{n \times n}$ is also a square matrix, so it makes sense to apply the trace operator to it. To verify that $\mathrm{tr}AB = \mathrm{tr}BA$, note that

$$\begin{aligned}
\mathrm{tr}AB &= \sum_{i=1}^{m} (AB)_{ii} = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} A_{ij} B_{ji} \right) \\
&= \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} B_{ji} = \sum_{j=1}^{n} \sum_{i=1}^{m} B_{ji} A_{ij} \\
&= \sum_{j=1}^{n} \left(\sum_{i=1}^{m} B_{ji} A_{ij} \right) = \sum_{j=1}^{n} (BA)_{jj} = \mathrm{tr}BA.
\end{aligned}$$

Here, the first and last two equalities use the definition of the trace operator and matrix multiplication. The fourth equality, where the main work occurs, uses the commutativity of scalar multiplication in order to reverse the order of the terms in each product, and the commutativity and associativity of scalar addition in order to rearrange the order of the summation.

3.5 Norms

A ***norm*** of a vector $\|x\|$ is informally a measure of the "length" of the vector. For example, we have the commonly-used Euclidean or ℓ_2 norm,

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

Note that $\|x\|_2^2 = x^T x$.

More formally, a norm is any function $f : \mathbb{R}^n \to \mathbb{R}$ that satisfies 4 properties:

1. For all $x \in \mathbb{R}^n$, $f(x) \geq 0$ (non-negativity).
2. $f(x) = 0$ if and only if $x = 0$ (definiteness).
3. For all $x \in \mathbb{R}^n$, $t \in \mathbb{R}$, $f(tx) = |t|f(x)$ (homogeneity).
4. For all $x, y \in \mathbb{R}^n$, $f(x + y) \leq f(x) + f(y)$ (triangle inequality).

Other examples of norms are the ℓ_1 norm,

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

and the ℓ_∞ norm,

$$\|x\|_\infty = \max_i |x_i|.$$

In fact, all three norms presented so far are examples of the family of ℓ_p norms, which are parameterized by a real number $p \geq 1$, and defined as

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

Norms can also be defined for matrices, such as the Frobenius norm,

$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2} = \sqrt{\text{tr}(A^T A)}.$$

Many other norms exist, but they are beyond the scope of this review.

3.6 Linear Independence and Rank

A set of vectors $\{x_1, x_2, \ldots x_n\} \subset \mathbb{R}^m$ is said to be ***(linearly) independent*** if no vector can be represented as a linear combination of the remaining vectors. Conversely, if one vector belonging to the set *can* be represented as a linear combination of the remaining vectors, then the vectors are said to be ***(linearly) dependent***. That is, if

$$x_n = \sum_{i=1}^{n-1} \alpha_i x_i$$

for some scalar values $\alpha_1, \ldots, \alpha_{n-1} \in \mathbb{R}$, then we say that the vectors $x_1, \ldots, x_n$ are linearly dependent; otherwise, the vectors are linearly independent. For example, the vectors

$$x_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad x_2 = \begin{bmatrix} 4 \\ 1 \\ 5 \end{bmatrix} \quad x_3 = \begin{bmatrix} 2 \\ -3 \\ -1 \end{bmatrix}$$

are linearly dependent because $x_3 = -2x_1 + x_2$.

The ***column rank*** of a matrix $A \in \mathbb{R}^{m \times n}$ is the size of the largest subset of columns of A that constitute a linearly independent set. With some abuse of terminology, this is often referred to simply as the number of linearly independent columns of A. In the same way, the ***row rank*** is the largest number of rows of A that constitute a linearly independent set.

For any matrix $A \in \mathbb{R}^{m \times n}$, it turns out that the column rank of A is equal to the row rank of A (though we will not prove this), and so both quantities are referred to collectively as the ***rank*** of A, denoted as $\text{rank}(A)$. The following are some basic properties of the rank:

- For $A \in \mathbb{R}^{m \times n}$, $\text{rank}(A) \leq \min(m, n)$. If $\text{rank}(A) = \min(m, n)$, then A is said to be ***full rank***.
- For $A \in \mathbb{R}^{m \times n}$, $\text{rank}(A) = \text{rank}(A^T)$.
- For $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$, $\text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B))$.
- For $A, B \in \mathbb{R}^{m \times n}$, $\text{rank}(A + B) \leq \text{rank}(A) + \text{rank}(B)$.

3.7 The Inverse

The ***inverse*** of a square matrix $A \in \mathbb{R}^{n \times n}$ is denoted A^{-1}, and is the unique matrix such that

$$A^{-1}A = I = AA^{-1}.$$

Note that not all matrices have inverses. Non-square matrices, for example, do not have inverses by definition. However, for some square matrices A, it may still be the case that

A^{-1} may not exist. In particular, we say that A is ***invertible*** or ***non-singular*** if A^{-1} exists and ***non-invertible*** or ***singular*** otherwise.[1]

In order for a square matrix A to have an inverse A^{-1}, then A must be full rank. We will soon see that there are many alternative sufficient and necessary conditions, in addition to full rank, for invertibility.

The following are properties of the inverse; all assume that $A, B \in \mathbb{R}^{n \times n}$ are non-singular:

- $(A^{-1})^{-1} = A$
- $(AB)^{-1} = B^{-1}A^{-1}$
- $(A^{-1})^T = (A^T)^{-1}$. For this reason this matrix is often denoted A^{-T}.

As an example of how the inverse is used, consider the linear system of equations, $Ax = b$ where $A \in \mathbb{R}^{n \times n}$, and $x, b \in \mathbb{R}^n$. If A is nonsingular (i.e., invertible), then $x = A^{-1}b$. (What if $A \in \mathbb{R}^{m \times n}$ is not a square matrix? Does this work?)

3.8 Orthogonal Matrices

Two vectors $x, y \in \mathbb{R}^n$ are ***orthogonal*** if $x^T y = 0$. A vector $x \in \mathbb{R}^n$ is ***normalized*** if $\|x\|_2 = 1$. A square matrix $U \in \mathbb{R}^{n \times n}$ is ***orthogonal*** (note the different meanings when talking about vectors versus matrices) if all its columns are orthogonal to each other and are normalized (the columns are then referred to as being ***orthonormal***).

It follows immediately from the definition of orthogonality and normality that

$$U^T U = I = U U^T.$$

In other words, the inverse of an orthogonal matrix is its transpose. Note that if U is not square — i.e., $U \in \mathbb{R}^{m \times n}$, $n < m$ — but its columns are still orthonormal, then $U^T U = I$, but $UU^T \neq I$. We generally only use the term orthogonal to describe the previous case, where U is square.

Another nice property of orthogonal matrices is that operating on a vector with an orthogonal matrix will not change its Euclidean norm, i.e.,

$$\|Ux\|_2 = \|x\|_2$$

for any $x \in \mathbb{R}^n$, $U \in \mathbb{R}^{n \times n}$ orthogonal.

3.9 Range and Nullspace of a Matrix

The ***span*** of a set of vectors $\{x_1, x_2, \ldots x_n\}$ is the set of all vectors that can be expressed as a linear combination of $\{x_1, \ldots, x_n\}$. That is,

$$\text{span}(\{x_1, \ldots x_n\}) = \left\{ v : v = \sum_{i=1}^{n} \alpha_i x_i, \;\; \alpha_i \in \mathbb{R} \right\}.$$

[1] It's easy to get confused and think that non-singular means non-invertible. But in fact, it means the opposite! Watch out!

It can be shown that if $\{x_1, \ldots, x_n\}$ is a set of n linearly independent vectors, where each $x_i \in \mathbb{R}^n$, then $\text{span}(\{x_1, \ldots x_n\}) = \mathbb{R}^n$. In other words, *any* vector $v \in \mathbb{R}^n$ can be written as a linear combination of x_1 through x_n. The ***projection*** of a vector $y \in \mathbb{R}^m$ onto the span of $\{x_1, \ldots, x_n\}$ (here we assume $x_i \in \mathbb{R}^m$) is the vector $v \in \text{span}(\{x_1, \ldots x_n\})$, such that v is as close as possible to y, as measured by the Euclidean norm $\|v - y\|_2$. We denote the projection as $\text{Proj}(y; \{x_1, \ldots, x_n\})$ and can define it formally as,

$$\text{Proj}(y; \{x_1, \ldots x_n\}) = \text{argmin}_{v \in \text{span}(\{x_1, \ldots, x_n\})} \|y - v\|_2.$$

The ***range*** (sometimes also called the columnspace) of a matrix $A \in \mathbb{R}^{m \times n}$, denoted $\mathcal{R}(A)$, is the the span of the columns of A. In other words,

$$\mathcal{R}(A) = \{v \in \mathbb{R}^m : v = Ax, x \in \mathbb{R}^n\}.$$

Making a few technical assumptions (namely that A is full rank and that $n < m$), the projection of a vector $y \in \mathbb{R}^m$ onto the range of A is given by,

$$\text{Proj}(y; A) = \text{argmin}_{v \in \mathcal{R}(A)} \|v - y\|_2 = A(A^T A)^{-1} A^T y \ .$$

This last equation should look extremely familiar, since it is almost the same formula we derived in class (and which we will soon derive again) for the least squares estimation of parameters. Looking at the definition for the projection, it should not be too hard to convince yourself that this is in fact the same objective that we minimized in our least squares problem (except for a squaring of the norm, which doesn't affect the optimal point) and so these problems are naturally very connected. When A contains only a single column, $a \in \mathbb{R}^m$, this gives the special case for a projection of a vector on to a line:

$$\text{Proj}(y; a) = \frac{aa^T}{a^T a} y \ .$$

The ***nullspace*** of a matrix $A \in \mathbb{R}^{m \times n}$, denoted $\mathcal{N}(A)$ is the set of all vectors that equal 0 when multiplied by A, i.e.,

$$\mathcal{N}(A) = \{x \in \mathbb{R}^n : Ax = 0\}.$$

Note that vectors in $\mathcal{R}(A)$ are of size m, while vectors in the $\mathcal{N}(A)$ are of size n, so vectors in $\mathcal{R}(A^T)$ and $\mathcal{N}(A)$ are both in $\mathbb{R}^n$. In fact, we can say much more. It turns out that

$$\left\{w : w = u + v, u \in \mathcal{R}(A^T), v \in \mathcal{N}(A)\right\} = \mathbb{R}^n \text{ and } \mathcal{R}(A^T) \cap \mathcal{N}(A) = \{\mathbf{0}\} \ .$$

In other words, $\mathcal{R}(A^T)$ and $\mathcal{N}(A)$ are disjoint subsets that together span the entire space of $\mathbb{R}^n$. Sets of this type are called ***orthogonal complements***, and we denote this $\mathcal{R}(A^T) = \mathcal{N}(A)^\perp$.

3.10 The Determinant

The ***determinant*** of a square matrix $A \in \mathbb{R}^{n \times n}$, is a function $\det : \mathbb{R}^{n \times n} \to \mathbb{R}$, and is denoted $|A|$ or $\det A$ (like the trace operator, we usually omit parentheses). Algebraically, one could write down an explicit formula for the determinant of A, but this unfortunately gives little intuition about its meaning. Instead, we'll start out by providing a geometric interpretation of the determinant and then visit some of its specific algebraic properties afterwards.

Given a matrix

$$\begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_n^T & — \end{bmatrix},$$

consider the set of points $S \subset \mathbb{R}^n$ formed by taking all possible linear combinations of the row vectors $a_1, \ldots, a_n \in \mathbb{R}^n$ of A, where the coefficients of the linear combination are all between 0 and 1; that is, the set S is the restriction of $\text{span}(\{a_1, \ldots, a_n\})$ to only those linear combinations whose coefficients $\alpha_1, \ldots, \alpha_n$ satisfy $0 \leq \alpha_i \leq 1$, $i = 1, \ldots, n$. Formally,

$$S = \{v \in \mathbb{R}^n : v = \sum_{i=1}^{n} \alpha_i a_i \text{ where } 0 \leq \alpha_i \leq 1, i = 1, \ldots, n\}.$$

The absolute value of the determinant of A, it turns out, is a measure of the "volume" of the set S.[2]

For example, consider the 2×2 matrix,

$$A = \begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix}. \tag{1}$$

Here, the rows of the matrix are

$$a_1 = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \quad a_2 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}.$$

The set S corresponding to these rows is shown in Figure 1. For two-dimensional matrices, S generally has the shape of a *parallelogram*. In our example, the value of the determinant is $|A| = -7$ (as can be computed using the formulas shown later in this section), so the area of the parallelogram is 7. (Verify this for yourself!)

In three dimensions, the set S corresponds to an object known as a *parallelepiped* (a three-dimensional box with skewed sides, such that every face has the shape of a parallelogram). The absolute value of the determinant of the 3×3 matrix whose rows define S give the three-dimensional volume of the parallelepiped. In even higher dimensions, the set S is an object known as an n-dimensional *parallelotope*.

[2]Admittedly, we have not actually defined what we mean by "volume" here, but hopefully the intuition should be clear enough. When $n = 2$, our notion of "volume" corresponds to the area of S in the Cartesian plane. When $n = 3$, "volume" corresponds with our usual notion of volume for a three-dimensional object.

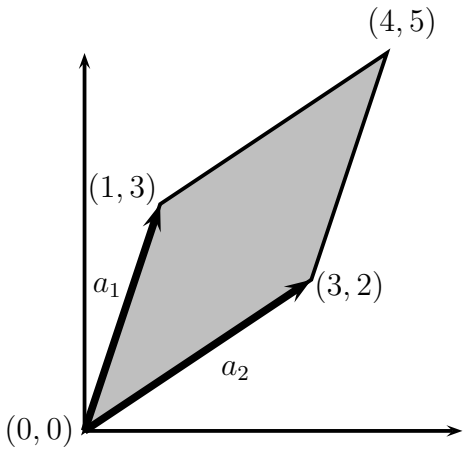

Figure 1: Illustration of the determinant for the 2×2 matrix A given in (1). Here, a_1 and a_2 are vectors corresponding to the rows of A, and the set S corresponds to the shaded region (i.e., the parallelogram). The absolute value of the determinant, $|\det A| = 7$, is the area of the parallelogram.

Algebraically, the determinant satisfies the following three properties (from which all other properties follow, including the general formula):

1. The determinant of the identity is 1, $|I| = 1$. (Geometrically, the volume of a unit hypercube is 1).

2. Given a matrix $A \in \mathbb{R}^{n \times n}$, if we multiply a single row in A by a scalar $t \in \mathbb{R}$, then the determinant of the new matrix is $t|A|$,

$$\left| \begin{bmatrix} — & t\, a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix} \right| = t|A|.$$

   (Geometrically, multiplying one of the sides of the set S by a factor t causes the volume to increase by a factor t.)

3. If we exchange any two rows a_i^T and a_j^T of A, then the determinant of the new matrix is $-|A|$, for example

$$\left| \begin{bmatrix} — & a_2^T & — \\ — & a_1^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix} \right| = -|A|.$$

In case you are wondering, it is not immediately obvious that a function satisfying the above three properties exists. In fact, though, such a function does exist, and is unique (which we will not prove here).

Several properties that follow from the three properties above include:

- For $A \in \mathbb{R}^{n\times n}$, $|A| = |A^T|$.
- For $A, B \in \mathbb{R}^{n\times n}$, $|AB| = |A||B|$.
- For $A \in \mathbb{R}^{n\times n}$, $|A| = 0$ if and only if A is singular (i.e., non-invertible). (If A is singular then it does not have full rank, and hence its columns are linearly dependent. In this case, the set S corresponds to a "flat sheet" within the n-dimensional space and hence has zero volume.)
- For $A \in \mathbb{R}^{n\times n}$ and A non-singular, $|A^{-1}| = 1/|A|$.

Before giving the general definition for the determinant, we define, for $A \in \mathbb{R}^{n\times n}$, $A_{\setminus i,\setminus j} \in \mathbb{R}^{(n-1)\times(n-1)}$ to be the *matrix* that results from deleting the ith row and jth column from A. The general (recursive) formula for the determinant is

$$\begin{aligned} |A| &= \sum_{i=1}^{n} (-1)^{i+j} a_{ij} |A_{\setminus i,\setminus j}| \quad \text{(for any } j \in 1,\ldots,n) \\ &= \sum_{j=1}^{n} (-1)^{i+j} a_{ij} |A_{\setminus i,\setminus j}| \quad \text{(for any } i \in 1,\ldots,n) \end{aligned}$$

with the initial case that $|A| = a_{11}$ for $A \in \mathbb{R}^{1\times 1}$. If we were to expand this formula completely for $A \in \mathbb{R}^{n\times n}$, there would be a total of $n!$ (n factorial) different terms. For this reason, we hardly ever explicitly write the complete equation of the determinant for matrices bigger than 3×3. However, the equations for determinants of matrices up to size 3×3 are fairly common, and it is good to know them:

$$\begin{aligned} |[a_{11}]| &= a_{11} \\ \left|\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\right| &= a_{11}a_{22} - a_{12}a_{21} \\ \left|\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}\right| &= \begin{array}{l} a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ \quad -a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31} \end{array} \end{aligned}$$

The ***classical adjoint*** (often just called the adjoint) of a matrix $A \in \mathbb{R}^{n\times n}$, is denoted $\text{adj}(A)$, and defined as

$$\text{adj}(A) \in \mathbb{R}^{n\times n}, \quad (\text{adj}(A))_{ij} = (-1)^{i+j} |A_{\setminus j,\setminus i}|$$

(note the switch in the indices $A_{\setminus j,\setminus i}$). It can be shown that for any nonsingular $A \in \mathbb{R}^{n\times n}$,

$$A^{-1} = \frac{1}{|A|}\text{adj}(A) \ .$$

While this is a nice "explicit" formula for the inverse of matrix, we should note that, numerically, there are in fact much more efficient ways of computing the inverse.

3.11 Quadratic Forms and Positive Semidefinite Matrices

Given a square matrix $A \in \mathbb{R}^{n \times n}$ and a vector $x \in \mathbb{R}^n$, the scalar value $x^T Ax$ is called a ***quadratic form***. Written explicitly, we see that

$$x^T Ax = \sum_{i=1}^{n} x_i (Ax)_i = \sum_{i=1}^{n} x_i \left(\sum_{j=1}^{n} A_{ij} x_j \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j \quad .$$

Note that,

$$x^T Ax = (x^T Ax)^T = x^T A^T x = x^T \left(\frac{1}{2} A + \frac{1}{2} A^T \right) x,$$

where the first equality follows from the fact that the transpose of a scalar is equal to itself, and the second equality follows from the fact that we are averaging two quantities which are themselves equal. From this, we can conclude that only the symmetric part of A contributes to the quadratic form. For this reason, we often implicitly assume that the matrices appearing in a quadratic form are symmetric.

We give the following definitions:

- A symmetric matrix $A \in \mathbb{S}^n$ is ***positive definite*** (PD) if for all non-zero vectors $x \in \mathbb{R}^n$, $x^T Ax > 0$. This is usually denoted $A \succ 0$ (or just $A > 0$), and often times the set of all positive definite matrices is denoted $\mathbb{S}^n_{++}$.
- A symmetric matrix $A \in \mathbb{S}^n$ is ***positive semidefinite*** (PSD) if for all vectors $x^T Ax \geq 0$. This is written $A \succeq 0$ (or just $A \geq 0$), and the set of all positive semidefinite matrices is often denoted $\mathbb{S}^n_{+}$.
- Likewise, a symmetric matrix $A \in \mathbb{S}^n$ is ***negative definite*** (ND), denoted $A \prec 0$ (or just $A < 0$) if for all non-zero $x \in \mathbb{R}^n$, $x^T Ax < 0$.
- Similarly, a symmetric matrix $A \in \mathbb{S}^n$ is ***negative semidefinite*** (NSD), denoted $A \preceq 0$ (or just $A \leq 0$) if for all $x \in \mathbb{R}^n$, $x^T Ax \leq 0$.
- Finally, a symmetric matrix $A \in \mathbb{S}^n$ is ***indefinite***, if it is neither positive semidefinite nor negative semidefinite — i.e., if there exists $x_1, x_2 \in \mathbb{R}^n$ such that $x_1^T Ax_1 > 0$ and $x_2^T Ax_2 < 0$.

It should be obvious that if A is positive definite, then $-A$ is negative definite and vice versa. Likewise, if A is positive semidefinite then $-A$ is negative semidefinite and vice versa. If A is indefinite, then so is $-A$.

One important property of positive definite and negative definite matrices is that they are always full rank, and hence, invertible. To see why this is the case, suppose that some matrix $A \in \mathbb{R}^{n \times n}$ is not full rank. Then, suppose that the jth column of A is expressible as a linear combination of other $n - 1$ columns:

$$a_j = \sum_{i \neq j} x_i a_i,$$

for some $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n \in \mathbb{R}$. Setting $x_j = -1$, we have

$$Ax = \sum_{i=1}^{n} x_i a_i = 0.$$

But this implies $x^T Ax = 0$ for some non-zero vector x, so A must be neither positive definite nor negative definite. Therefore, if A is either positive definite or negative definite, it must be full rank.

Finally, there is one type of positive definite matrix that comes up frequently, and so deserves some special mention. Given any matrix $A \in \mathbb{R}^{m \times n}$ (not necessarily symmetric or even square), the matrix $G = A^T A$ (sometimes called a ***Gram matrix***) is always positive semidefinite. Further, if $m \geq n$ (and we assume for convenience that A is full rank), then $G = A^T A$ is positive definite.

3.12 Eigenvalues and Eigenvectors

Given a square matrix $A \in \mathbb{R}^{n \times n}$, we say that $\lambda \in \mathbb{C}$ is an ***eigenvalue*** of A and $x \in \mathbb{C}^n$ is the corresponding ***eigenvector***[3] if

$$Ax = \lambda x, \quad x \neq 0.$$

Intuitively, this definition means that multiplying A by the vector x results in a new vector that points in the same direction as x, but scaled by a factor λ. Also note that for any eigenvector $x \in \mathbb{C}^n$, and scalar $t \in \mathbb{C}$, $A(cx) = cAx = c\lambda x = \lambda(cx)$, so cx is also an eigenvector. For this reason when we talk about "the" eigenvector associated with λ, we usually assume that the eigenvector is normalized to have length 1 (this still creates some ambiguity, since x and $-x$ will both be eigenvectors, but we will have to live with this).

We can rewrite the equation above to state that (λ, x) is an eigenvalue-eigenvector pair of A if,

$$(\lambda I - A)x = 0, \quad x \neq 0.$$

But $(\lambda I - A)x = 0$ has a non-zero solution to x if and only if $(\lambda I - A)$ has a non-empty nullspace, which is only the case if $(\lambda I - A)$ is singular, i.e.,

$$|(\lambda I - A)| = 0.$$

We can now use the previous definition of the determinant to expand this expression into a (very large) polynomial in λ, where λ will have maximum degree n. We then find the n (possibly complex) roots of this polynomial to find the n eigenvalues $\lambda_1, \ldots, \lambda_n$. To find the eigenvector corresponding to the eigenvalue λ_i, we simply solve the linear equation $(\lambda_i I - A)x = 0$. It should be noted that this is not the method which is actually used

[3]Note that λ and the entries of x are actually in $\mathbb{C}$, the set of complex numbers, not just the reals; we will see shortly why this is necessary. Don't worry about this technicality for now, you can think of complex vectors in the same way as real vectors.

in practice to numerically compute the eigenvalues and eigenvectors (remember that the complete expansion of the determinant has $n!$ terms); it is rather a mathematical argument.

The following are properties of eigenvalues and eigenvectors (in all cases assume $A \in \mathbb{R}^{n \times n}$ has eigenvalues $\lambda_i, \ldots, \lambda_n$ and associated eigenvectors $x_1, \ldots x_n$):

- The trace of a A is equal to the sum of its eigenvalues,

$$\mathrm{tr}A = \sum_{i=1}^{n} \lambda_i.$$

- The determinant of A is equal to the product of its eigenvalues,

$$|A| = \prod_{i=1}^{n} \lambda_i.$$

- The rank of A is equal to the number of non-zero eigenvalues of A.
- If A is non-singular then $1/\lambda_i$ is an eigenvalue of A^{-1} with associated eigenvector x_i, i.e., $A^{-1}x_i = (1/\lambda_i)x_i$. (To prove this, take the eigenvector equation, $Ax_i = \lambda_i x_i$ and left-multiply each side by A^{-1}.)
- The eigenvalues of a diagonal matrix $D = \mathrm{diag}(d_1, \ldots d_n)$ are just the diagonal entries $d_1, \ldots d_n$.

We can write all the eigenvector equations simultaneously as

$$AX = X\Lambda$$

where the columns of $X \in \mathbb{R}^{n \times n}$ are the eigenvectors of A and Λ is a diagonal matrix whose entries are the eigenvalues of A, i.e.,

$$X \in \mathbb{R}^{n \times n} = \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{bmatrix}, \quad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n).$$

If the eigenvectors of A are linearly independent, then the matrix X will be invertible, so $A = X\Lambda X^{-1}$. A matrix that can be written in this form is called ***diagonalizable***.

3.13 Eigenvalues and Eigenvectors of Symmetric Matrices

Two remarkable properties come about when we look at the eigenvalues and eigenvectors of a symmetric matrix $A \in \mathbb{S}^n$. First, it can be shown that all the eigenvalues of A are real. Secondly, the eigenvectors of A are orthonormal, i.e., the matrix X defined above is an orthogonal matrix (for this reason, we denote the matrix of eigenvectors as U in this case).

We can therefore represent A as $A = U\Lambda U^T$, remembering from above that the inverse of an orthogonal matrix is just its transpose.

Using this, we can show that the definiteness of a matrix depends entirely on the sign of its eigenvalues. Suppose $A \in \mathbb{S}^n = U\Lambda U^T$. Then

$$x^T Ax = x^T U\Lambda U^T x = y^T \Lambda y = \sum_{i=1}^{n} \lambda_i y_i^2$$

where $y = U^T x$ (and since U is full rank, any vector $y \in \mathbb{R}^n$ can be represented in this form). Because y_i^2 is always positive, the sign of this expression depends entirely on the λ_i's. If all $\lambda_i > 0$, then the matrix is positive definite; if all $\lambda_i \geq 0$, it is positive semidefinite. Likewise, if all $\lambda_i < 0$ or $\lambda_i \leq 0$, then A is negative definite or negative semidefinite respectively. Finally, if A has both positive and negative eigenvalues, it is indefinite.

An application where eigenvalues and eigenvectors come up frequently is in maximizing some function of a matrix. In particular, for a matrix $A \in \mathbb{S}^n$, consider the following maximization problem,

$$\max_{x \in \mathbb{R}^n} \; x^T Ax \quad \text{subject to } \|x\|_2^2 = 1$$

i.e., we want to find the vector (of norm 1) which maximizes the quadratic form. Assuming the eigenvalues are ordered as $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$, the optimal x for this optimization problem is x_1, the eigenvector corresponding to λ_1. In this case the maximal value of the quadratic form is λ_1. Similarly, the optimal solution to the minimization problem,

$$\min_{x \in \mathbb{R}^n} \; x^T Ax \quad \text{subject to } \|x\|_2^2 = 1$$

is x_n, the eigenvector corresponding to λ_n, and the minimal value is λ_n. This can be proved by appealing to the eigenvector-eigenvalue form of A and the properties of orthogonal matrices. However, in the next section we will see a way of showing it directly using matrix calculus.

4 Matrix Calculus

While the topics in the previous sections are typically covered in a standard course on linear algebra, one topic that does not seem to be covered very often (and which we will use extensively) is the extension of calculus to the vector setting. Despite the fact that all the actual calculus we use is relatively trivial, the notation can often make things look much more difficult than they are. In this section we present some basic definitions of matrix calculus and provide a few examples.

4.1 The Gradient

Suppose that $f : \mathbb{R}^{m \times n} \to \mathbb{R}$ is a function that takes as input a matrix A of size $m \times n$ and returns a real value. Then the ***gradient*** of f (with respect to $A \in \mathbb{R}^{m \times n}$) is the matrix of

partial derivatives, defined as:

$$\nabla_A f(A) \in \mathbb{R}^{m \times n} = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{12}} & \cdots & \frac{\partial f(A)}{\partial A_{1n}} \\ \frac{\partial f(A)}{\partial A_{21}} & \frac{\partial f(A)}{\partial A_{22}} & \cdots & \frac{\partial f(A)}{\partial A_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{m1}} & \frac{\partial f(A)}{\partial A_{m2}} & \cdots & \frac{\partial f(A)}{\partial A_{mn}} \end{bmatrix}$$

i.e., an $m \times n$ matrix with

$$(\nabla_A f(A))_{ij} = \frac{\partial f(A)}{\partial A_{ij}}.$$

Note that the size of $\nabla_A f(A)$ is always the same as the size of A. So if, in particular, A is just a vector $x \in \mathbb{R}^n$,

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}.$$

It is very important to remember that the gradient of a function is *only* defined if the function is real-valued, that is, if it returns a scalar value. We can not, for example, take the gradient of $Ax, A \in \mathbb{R}^{n \times n}$ with respect to x, since this quantity is vector-valued.

It follows directly from the equivalent properties of partial derivatives that:

- $\nabla_x(f(x) + g(x)) = \nabla_x f(x) + \nabla_x g(x)$.
- For $t \in \mathbb{R}$, $\nabla_x(t\ f(x)) = t\nabla_x f(x)$.

In principle, gradients are a natural extension of partial derivatives to functions of multiple variables. In practice, however, working with gradients can sometimes be tricky for notational reasons. For example, suppose that $A \in \mathbb{R}^{m \times n}$ is a matrix of fixed coefficients and suppose that $b \in \mathbb{R}^m$ is a vector of fixed coefficients. Let $f : \mathbb{R}^m \to \mathbb{R}$ be the function defined by $f(z) = z^T z$, such that $\nabla_z f(z) = 2z$. But now, consider the expression,

$$\nabla f(Ax).$$

How should this expression be interpreted? There are at least two possibilities:

1. In the first interpretation, recall that $\nabla_z f(z) = 2z$. Here, we interpret $\nabla f(Ax)$ as evaluating the gradient at the point Ax, hence,

$$\nabla f(Ax) = 2(Ax) = 2Ax \in \mathbb{R}^m.$$

2. In the second interpretation, we consider the quantity $f(Ax)$ as a function of the input variables x. More formally, let $g(x) = f(Ax)$. Then in this interpretation,

$$\nabla f(Ax) = \nabla_x g(x) \in \mathbb{R}^n.$$

Here, we can see that these two interpretations are indeed different. One interpretation yields an m-dimensional vector as a result, while the other interpretation yields an n-dimensional vector as a result! How can we resolve this?

Here, the key is to make explicit the variables which we are differentiating with respect to. In the first case, we are differentiating the function f with respect to its arguments z and then substituting the argument Ax. In the second case, we are differentiating the composite function $g(x) = f(Ax)$ with respect to x directly. We denote the first case as $\nabla_z f(Ax)$ and the second case as $\nabla_x f(Ax)$.[4] Keeping the notation clear is extremely important (as you'll find out in your homework, in fact!).

4.2 The Hessian

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a function that takes a vector in $\mathbb{R}^n$ and returns a real number. Then the ***Hessian*** matrix with respect to x, written $\nabla_x^2 f(x)$ or simply as H is the $n \times n$ matrix of partial derivatives,

$$\nabla_x^2 f(x) \in \mathbb{R}^{n \times n} = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}.$$

In other words, $\nabla_x^2 f(x) \in \mathbb{R}^{n \times n}$, with

$$(\nabla_x^2 f(x))_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}.$$

Note that the Hessian is always symmetric, since

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x)}{\partial x_j \partial x_i}.$$

Similar to the gradient, the Hessian is defined only when $f(x)$ is real-valued.

It is natural to think of the gradient as the analogue of the first derivative for functions of vectors, and the Hessian as the analogue of the second derivative (and the symbols we use also suggest this relation). This intuition is generally correct, but there a few caveats to keep in mind.

[4] A drawback to this notation that we will have to live with is the fact that in the first case, $\nabla_z f(Ax)$ it appears that we are differentiating with respect to a variable that does not even appear in the expression being differentiated! For this reason, the first case is often written as $\nabla f(Ax)$, and the fact that we are differentiating with respect to the arguments of f is understood. However, the second case is *always* written as $\nabla_x f(Ax)$.

First, for real-valued functions of one variable $f : \mathbb{R} \to \mathbb{R}$, it is a basic definition that the second derivative is the derivative of the first derivative, i.e.,

$$\frac{\partial^2 f(x)}{\partial x^2} = \frac{\partial}{\partial x}\frac{\partial}{\partial x} f(x).$$

However, for functions of a vector, the gradient of the function is a vector, and we cannot take the gradient of a vector — i.e.,

$$\nabla_x \nabla_x f(x) = \nabla_x \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

and this expression is not defined. Therefore, it is *not* the case that the Hessian is the gradient of the gradient. However, this is *almost* true, in the following sense: If we look at the ith entry of the gradient $(\nabla_x f(x))_i = \partial f(x)/\partial x_i$, and take the gradient with respect to x we get

$$\nabla_x \frac{\partial f(x)}{\partial x_i} = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_i \partial x_1} \\ \frac{\partial^2 f(x)}{\partial x_i \partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_i \partial x_n} \end{bmatrix}$$

which is the ith column (or row) of the Hessian. Therefore,

$$\nabla_x^2 f(x) = \begin{bmatrix} \nabla_x(\nabla_x f(x))_1 & \nabla_x(\nabla_x f(x))_2 & \cdots & \nabla_x(\nabla_x f(x))_n \end{bmatrix}.$$

If we don't mind being a little bit sloppy we can say that (essentially) $\nabla_x^2 f(x) = \nabla_x(\nabla_x f(x))^T$, so long as we understand that this really means taking the gradient of each entry of $(\nabla_x f(x))^T$, not the gradient of the whole vector.

Finally, note that while we can take the gradient with respect to a matrix $A \in \mathbb{R}^n$, for the purposes of this class we will only consider taking the Hessian with respect to a vector $x \in \mathbb{R}^n$. This is simply a matter of convenience (and the fact that none of the calculations we do require us to find the Hessian with respect to a matrix), since the Hessian with respect to a matrix would have to represent all the partial derivatives $\partial^2 f(A)/(\partial A_{ij} \partial A_{k\ell})$, and it is rather cumbersome to represent this as a matrix.

4.3 Gradients and Hessians of Quadratic and Linear Functions

Now let's try to determine the gradient and Hessian matrices for a few simple functions. It should be noted that all the gradients given here are special cases of the gradients given in the CS229 lecture notes.

For $x \in \mathbb{R}^n$, let $f(x) = b^T x$ for some known vector $b \in \mathbb{R}^n$. Then

$$f(x) = \sum_{i=1}^{n} b_i x_i$$

so

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} b_i x_i = b_k.$$

From this we can easily see that $\nabla_x b^T x = b$. This should be compared to the analogous situation in single variable calculus, where $\partial/(\partial x)\ ax = a$.

Now consider the quadratic function $f(x) = x^T A x$ for $A \in \mathbb{S}^n$. Remember that

$$f(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j.$$

To take the partial derivative, we'll consider the terms including x_k and x_k^2 factors separately:

$$\begin{aligned}
\frac{\partial f(x)}{\partial x_k} &= \frac{\partial}{\partial x_k} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j \\
&= \frac{\partial}{\partial x_k} \left[\sum_{i \neq k} \sum_{j \neq k} A_{ij} x_i x_j + \sum_{i \neq k} A_{ik} x_i x_k + \sum_{j \neq k} A_{kj} x_k x_j + A_{kk} x_k^2 \right] \\
&= \sum_{i \neq k} A_{ik} x_i + \sum_{j \neq k} A_{kj} x_j + 2 A_{kk} x_k \\
&= \sum_{i=1}^{n} A_{ik} x_i + \sum_{j=1}^{n} A_{kj} x_j = 2 \sum_{i=1}^{n} A_{ki} x_i,
\end{aligned}$$

where the last equality follows since A is symmetric (which we can safely assume, since it is appearing in a quadratic form). Note that the kth entry of $\nabla_x f(x)$ is just the inner product of the kth row of A and x. Therefore, $\nabla_x x^T A x = 2Ax$. Again, this should remind you of the analogous fact in single-variable calculus, that $\partial/(\partial x)\ ax^2 = 2ax$.

Finally, let's look at the Hessian of the quadratic function $f(x) = x^T A x$ (it should be obvious that the Hessian of a linear function $b^T x$ is zero). In this case,

$$\frac{\partial^2 f(x)}{\partial x_k \partial x_\ell} = \frac{\partial}{\partial x_k} \left[\frac{\partial f(x)}{\partial x_\ell} \right] = \frac{\partial}{\partial x_k} \left[2 \sum_{i=1}^{n} A_{\ell i} x_i \right] = 2A_{\ell k} = 2A_{k\ell}.$$

Therefore, it should be clear that $\nabla_x^2 x^T A x = 2A$, which should be entirely expected (and again analogous to the single-variable fact that $\partial^2/(\partial x^2)\ ax^2 = 2a$).

To recap,

- $\nabla_x b^T x = b$
- $\nabla_x x^T A x = 2Ax$ (if A symmetric)
- $\nabla_x^2 x^T A x = 2A$ (if A symmetric)

4.4 Least Squares

Let's apply the equations we obtained in the last section to derive the least squares equations. Suppose we are given matrices $A \in \mathbb{R}^{m\times n}$ (for simplicity we assume A is full rank) and a vector $b \in \mathbb{R}^m$ such that $b \notin \mathcal{R}(A)$. In this situation we will not be able to find a vector $x \in \mathbb{R}^n$, such that $Ax = b$, so instead we want to find a vector x such that Ax is as close as possible to b, as measured by the square of the Euclidean norm $\|Ax - b\|_2^2$.

Using the fact that $\|x\|_2^2 = x^T x$, we have

$$\begin{aligned} \|Ax - b\|_2^2 &= (Ax - b)^T (Ax - b) \\ &= x^T A^T Ax - 2b^T Ax + b^T b \end{aligned}$$

Taking the gradient with respect to x we have, and using the properties we derived in the previous section

$$\begin{aligned} \nabla_x (x^T A^T Ax - 2b^T Ax + b^T b) &= \nabla_x x^T A^T Ax - \nabla_x 2b^T Ax + \nabla_x b^T b \\ &= 2A^T Ax - 2A^T b \end{aligned}$$

Setting this last expression equal to zero and solving for x gives the normal equations

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

which is the same as what we derived in class.

4.5 Gradients of the Determinant

Now let's consider a situation where we find the gradient of a function with respect to a matrix, namely for $A \in \mathbb{R}^{n\times n}$, we want to find $\nabla_A |A|$. Recall from our discussion of determinants that

$$|A| = \sum_{i=1}^{n} (-1)^{i+j} A_{ij} |A_{\setminus i, \setminus j}| \quad \text{(for any } j \in 1, \ldots, n)$$

so

$$\frac{\partial}{\partial A_{k\ell}} |A| = \frac{\partial}{\partial A_{k\ell}} \sum_{i=1}^{n} (-1)^{i+j} A_{ij} |A_{\setminus i, \setminus j}| = (-1)^{k+\ell} |A_{\setminus k, \setminus \ell}| = (\text{adj}(A))_{\ell k}.$$

From this it immediately follows from the properties of the adjoint that

$$\nabla_A |A| = (\text{adj}(A))^T = |A| A^{-T}.$$

Now let's consider the function $f : \mathbb{S}_{++}^n \to \mathbb{R}$, $f(A) = \log |A|$. Note that we have to restrict the domain of f to be the positive definite matrices, since this ensures that $|A| > 0$, so that the log of $|A|$ is a real number. In this case we can use the chain rule (nothing fancy, just the ordinary chain rule from single-variable calculus) to see that

$$\frac{\partial \log |A|}{\partial A_{ij}} = \frac{\partial \log |A|}{\partial |A|} \frac{\partial |A|}{\partial A_{ij}} = \frac{1}{|A|} \frac{\partial |A|}{\partial A_{ij}}.$$

From this it should be obvious that

$$\nabla_A \log |A| = \frac{1}{|A|} \nabla_A |A| = A^{-1},$$

where we can drop the transpose in the last expression because A is symmetric. Note the similarity to the single-valued case, where $\partial/(\partial x)\ \log x = 1/x$.

4.6 Eigenvalues as Optimization

Finally, we use matrix calculus to solve an optimization problem in a way that leads directly to eigenvalue/eigenvector analysis. Consider the following, equality constrained optimization problem:

$$\max_{x \in \mathbb{R}^n}\ x^T A x \quad \text{subject to } \|x\|_2^2 = 1$$

for a symmetric matrix $A \in \mathbb{S}^n$. A standard way of solving optimization problems with equality constraints is by forming the ***Lagrangian***, an objective function that includes the equality constraints.[5] The Lagrangian in this case can be given by

$$\mathcal{L}(x, \lambda) = x^T A x - \lambda x^T x$$

where λ is called the Lagrange multiplier associated with the equality constraint. It can be established that for x^* to be a optimal point to the problem, the gradient of the Lagrangian has to be zero at x^* (this is not the only condition, but it is required). That is,

$$\nabla_x \mathcal{L}(x, \lambda) = \nabla_x (x^T A x - \lambda x^T x) = 2A^T x - 2\lambda x = 0.$$

Notice that this is just the linear equation $Ax = \lambda x$. This shows that the only points which can possibly maximize (or minimize) $x^T A x$ assuming $x^T x = 1$ are the eigenvectors of A.

[5] Don't worry if you haven't seen Lagrangians before, as we will cover them in greater detail later in CS229.

Linear algebra explained in four pages

Excerpt from the NO BULLSHIT GUIDE TO LINEAR ALGEBRA by Ivan Savov

***Abstract*—This document will review the fundamental ideas of linear algebra. We will learn about matrices, matrix operations, linear transformations and discuss both the theoretical and computational aspects of linear algebra. The tools of linear algebra open the gateway to the study of more advanced mathematics. A lot of *knowledge buzz* awaits you if you choose to follow the path of *understanding*, instead of trying to memorize a bunch of formulas.**

I. Introduction

Linear algebra is the math of vectors and matrices. Let n be a positive integer and let $\mathbb{R}$ denote the set of real numbers, then $\mathbb{R}^n$ is the set of all n-tuples of real numbers. A vector $\vec{v} \in \mathbb{R}^n$ is an n-tuple of real numbers. The notation "$\in S$" is read "element of S." For example, consider a vector that has three components:

$$\vec{v} = (v_1, v_2, v_3) \ \in \ (\mathbb{R}, \mathbb{R}, \mathbb{R}) \equiv \mathbb{R}^3.$$

A matrix $A \in \mathbb{R}^{m \times n}$ is a rectangular array of real numbers with m rows and n columns. For example, a 3×2 matrix looks like this:

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \in \begin{bmatrix} \mathbb{R} & \mathbb{R} \\ \mathbb{R} & \mathbb{R} \\ \mathbb{R} & \mathbb{R} \end{bmatrix} \equiv \mathbb{R}^{3\times 2}.$$

The purpose of this document is to introduce you to the mathematical operations that we can perform on vectors and matrices and to give you a feel of the power of linear algebra. Many problems in science, business, and technology can be described in terms of vectors and matrices so it is important that you understand how to work with these.

Prerequisites

The only prerequisite for this tutorial is a basic understanding of high school math concepts[1] like numbers, variables, equations, and the fundamental arithmetic operations on real numbers: addition (denoted $+$), subtraction (denoted $-$), multiplication (denoted implicitly), and division (fractions).

You should also be familiar with *functions* that take real numbers as inputs and give real numbers as outputs, $f : \mathbb{R} \to \mathbb{R}$. Recall that, by definition, the *inverse function* f^{-1} *undoes* the effect of f. If you are given $f(x)$ and you want to find x, you can use the inverse function as follows: $f^{-1}(f(x)) = x$. For example, the function $f(x) = \ln(x)$ has the inverse $f^{-1}(x) = e^x$, and the inverse of $g(x) = \sqrt{x}$ is $g^{-1}(x) = x^2$.

II. Definitions

A. Vector operations

We now define the math operations for vectors. The operations we can perform on vectors $\vec{u} = (u_1, u_2, u_3)$ and $\vec{v} = (v_1, v_2, v_3)$ are: addition, subtraction, scaling, norm (length), dot product, and cross product:

$$\begin{aligned} \vec{u} + \vec{v} &= (u_1 + v_1, u_2 + v_2, u_3 + v_3) \\ \vec{u} - \vec{v} &= (u_1 - v_1, u_2 - v_2, u_3 - v_3) \\ \alpha\vec{u} &= (\alpha u_1, \alpha u_2, \alpha u_3) \\ ||\vec{u}|| &= \sqrt{u_1^2 + u_2^2 + u_3^2} \\ \vec{u} \cdot \vec{v} &= u_1 v_1 + u_2 v_2 + u_3 v_3 \\ \vec{u} \times \vec{v} &= (u_2 v_3 - u_3 v_2, \ u_3 v_1 - u_1 v_3, \ u_1 v_2 - u_2 v_1) \end{aligned}$$

The dot product and the cross product of two vectors can also be described in terms of the angle θ between the two vectors. The formula for the dot product of the vectors is $\vec{u} \cdot \vec{v} = \|\vec{u}\|\|\vec{v}\| \cos\theta$. We say two vectors $\vec{u}$ and $\vec{v}$ are *orthogonal* if the angle between them is 90°. The dot product of orthogonal vectors is zero: $\vec{u} \cdot \vec{v} = \|\vec{u}\|\|\vec{v}\| \cos(90^\circ) = 0$.

The *norm* of the cross product is given by $\|\vec{u} \times \vec{v}\| = \|\vec{u}\|\|\vec{v}\| \sin\theta$. The cross product is not commutative: $\vec{u} \times \vec{v} \neq \vec{v} \times \vec{u}$, in fact $\vec{u} \times \vec{v} = -\vec{v} \times \vec{u}$.

[1] A good textbook to (re)learn high school math is minireference.com

B. Matrix operations

We denote by A the matrix as a whole and refer to its entries as a_{ij}. The mathematical operations defined for matrices are the following:

- addition (denoted $+$)

$$C = A + B \qquad \Leftrightarrow \qquad c_{ij} = a_{ij} + b_{ij}.$$

- subtraction (the inverse of addition)
- matrix product. The product of matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times \ell}$ is another matrix $C \in \mathbb{R}^{m\times \ell}$ given by the formula

$$C = AB \qquad \Leftrightarrow \qquad c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj},$$

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \\ a_{31}b_{11} + a_{32}b_{21} & a_{31}b_{12} + a_{32}b_{22} \end{bmatrix}$$

- matrix inverse (denoted A^{-1})
- matrix transpose (denoted $^\mathsf{T}$):

$$\begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 \\ \beta_1 & \beta_2 & \beta_3 \end{bmatrix}^\mathsf{T} = \begin{bmatrix} \alpha_1 & \beta_1 \\ \alpha_2 & \beta_2 \\ \alpha_3 & \beta_3 \end{bmatrix}.$$

- matrix trace: $\text{Tr}[A] \equiv \sum_{i=1}^{n} a_{ii}$
- determinant (denoted $\det(A)$ or $|A|$)

Note that the matrix product is not a commutative operation: $AB \neq BA$.

C. Matrix-vector product

The matrix-vector product is an important special case of the matrix-matrix product. The product of a 3×2 matrix A and the 2×1 column vector $\vec{x}$ results in a 3×1 vector $\vec{y}$ given by:

$$\vec{y} = A\vec{x} \qquad \Leftrightarrow \qquad \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \\ a_{31}x_1 + a_{32}x_2 \end{bmatrix}$$

$$= x_1 \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \\ a_{32} \end{bmatrix} \qquad \text{(C)}$$

$$= \begin{bmatrix} (a_{11}, a_{12}) \cdot \vec{x} \\ (a_{21}, a_{22}) \cdot \vec{x} \\ (a_{31}, a_{32}) \cdot \vec{x} \end{bmatrix}. \qquad \text{(R)}$$

There are two[2] fundamentally different yet equivalent ways to interpret the matrix-vector product. In the column picture, **(C)**, the multiplication of the matrix A by the vector $\vec{x}$ produces a **linear combination of the columns of the matrix**: $\vec{y} = A\vec{x} = x_1 A_{[:,1]} + x_2 A_{[:,2]}$, where $A_{[:,1]}$ and $A_{[:,2]}$ are the first and second columns of the matrix A.

In the row picture, **(R)**, multiplication of the matrix A by the vector $\vec{x}$ produces a column vector with coefficients equal to the **dot products of rows of the matrix** with the vector $\vec{x}$.

D. Linear transformations

The matrix-vector product is used to define the notion of a *linear transformation*, which is one of the key notions in the study of linear algebra. Multiplication by a matrix $A \in \mathbb{R}^{m\times n}$ can be thought of as computing a *linear transformation* T_A that takes n-vectors as inputs and produces m-vectors as outputs:

$$T_A : \mathbb{R}^n \to \mathbb{R}^m.$$

[2] For more info see the video of Prof. Strang's MIT lecture: bit.ly/10vmKcL

Instead of writing $\vec{y} = T_A(\vec{x})$ for the linear transformation T_A applied to the vector $\vec{x}$, we simply write $\vec{y} = A\vec{x}$. Applying the linear transformation T_A to the vector $\vec{x}$ corresponds to the product of the matrix A and the column vector $\vec{x}$. We say T_A is *represented by* the matrix A.

You can think of linear transformations as "vector functions" and describe their properties in analogy with the regular functions you are familiar with:

function $f : \mathbb{R} \to \mathbb{R}$	$\Leftrightarrow$	linear transformation $T_A : \mathbb{R}^n \to \mathbb{R}^m$
input $x \in \mathbb{R}$	$\Leftrightarrow$	input $\vec{x} \in \mathbb{R}^n$
output $f(x)$	$\Leftrightarrow$	output $T_A(\vec{x}) = A\vec{x} \in \mathbb{R}^m$
$g \circ f = g(f(x))$	$\Leftrightarrow$	$T_B(T_A(\vec{x})) = BA\vec{x}$
function inverse f^{-1}	$\Leftrightarrow$	matrix inverse A^{-1}
zeros of f	$\Leftrightarrow$	$\mathcal{N}(A) \equiv$ null space of A
range of f	$\Leftrightarrow$	$\mathcal{C}(A) \equiv$ column space of A = range of T_A

Note that the combined effect of applying the transformation T_A followed by T_B on the input vector $\vec{x}$ is equivalent to the matrix product $BA\vec{x}$.

E. Fundamental vector spaces

A *vector space* consists of a set of vectors and all linear combinations of these vectors. For example the vector space $\mathcal{S} = \text{span}\{\vec{v}_1, \vec{v}_2\}$ consists of all vectors of the form $\vec{v} = \alpha\vec{v}_1 + \beta\vec{v}_2$, where α and β are real numbers. We now define three fundamental vector spaces associated with a matrix A.

The *column space* of a matrix A is the set of vectors that can be produced as linear combinations of the columns of the matrix A:

$$\mathcal{C}(A) \equiv \{\vec{y} \in \mathbb{R}^m \mid \vec{y} = A\vec{x} \text{ for some } \vec{x} \in \mathbb{R}^n\}.$$

The column space is the *range* of the linear transformation T_A (the set of possible outputs). You can convince yourself of this fact by reviewing the definition of the matrix-vector product in the column picture **(C)**. The vector $A\vec{x}$ contains x_1 times the 1st column of A, x_2 times the 2nd column of A, etc. Varying over all possible inputs $\vec{x}$, we obtain all possible linear combinations of the columns of A, hence the name "column space."

The *null space* $\mathcal{N}(A)$ of a matrix $A \in \mathbb{R}^{m\times n}$ consists of all the vectors that the matrix A sends to the zero vector:

$$\mathcal{N}(A) \equiv \{\vec{x} \in \mathbb{R}^n \mid A\vec{x} = \vec{0}\}.$$

The vectors in the null space are *orthogonal* to all the rows of the matrix. We can see this from the row picture **(R)**: the output vectors is $\vec{0}$ if and only if the input vector $\vec{x}$ is orthogonal to all the rows of A.

The *row space* of a matrix A, denoted $\mathcal{R}(A)$, is the set of linear combinations of the rows of A. The row space $\mathcal{R}(A)$ is the orthogonal complement of the null space $\mathcal{N}(A)$. This means that for all vectors $\vec{v} \in \mathcal{R}(A)$ and all vectors $\vec{w} \in \mathcal{N}(A)$, we have $\vec{v} \cdot \vec{w} = 0$. Together, the null space and the row space form the domain of the transformation T_A, $\mathbb{R}^n = \mathcal{N}(A) \oplus \mathcal{R}(A)$, where $\oplus$ stands for *orthogonal direct sum*.

F. Matrix inverse

By definition, the inverse matrix A^{-1} *undoes* the effects of the matrix A. The cumulative effect of applying A^{-1} after A is the identity matrix $\mathbb{1}$:

$$A^{-1}A = \mathbb{1} \equiv \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}.$$

The identity matrix (ones on the diagonal and zeros everywhere else) corresponds to the identity transformation: $T_{\mathbb{1}}(\vec{x}) = \mathbb{1}\vec{x} = \vec{x}$, for all $\vec{x}$.

The matrix inverse is useful for solving matrix equations. Whenever we want to get rid of the matrix A in some matrix equation, we can "hit" A with its inverse A^{-1} to make it disappear. For example, to solve for the matrix X in the equation $XA = B$, multiply both sides of the equation by A^{-1} from the right: $X = BA^{-1}$. To solve for X in $ABCXD = E$, multiply both sides of the equation by D^{-1} on the right and by A^{-1}, B^{-1} and C^{-1} (in that order) from the left: $X = C^{-1}B^{-1}A^{-1}ED^{-1}$.

III. Computational linear algebra

Okay, I hear what you are saying "Dude, enough with the theory talk, let's see some calculations." In this section we'll look at one of the fundamental algorithms of linear algebra called Gauss–Jordan elimination.

A. Solving systems of equations

Suppose we're asked to solve the following system of equations:

$$\begin{aligned} 1x_1 + 2x_2 &= 5, \\ 3x_1 + 9x_2 &= 21. \end{aligned} \tag{1}$$

Without a knowledge of linear algebra, we could use substitution, elimination, or subtraction to find the values of the two unknowns x_1 and x_2.

Gauss–Jordan elimination is a systematic procedure for solving systems of equations based the following *row operations*:

α) Adding a multiple of one row to another row
β) Swapping two rows
γ) Multiplying a row by a constant

These row operations allow us to simplify the system of equations without changing their solution.

To illustrate the Gauss–Jordan elimination procedure, we'll now show the sequence of row operations required to solve the system of linear equations described above. We start by constructing an *augmented matrix* as follows:

$$\left[\begin{array}{cc|c} \mathbf{1} & 2 & 5 \\ 3 & 9 & 21 \end{array}\right].$$

The first column in the augmented matrix corresponds to the coefficients of the variable x_1, the second column corresponds to the coefficients of x_2, and the third column contains the constants from the right-hand side.

The Gauss-Jordan elimination procedure consists of two phases. During the first phase, we proceed left-to-right by choosing a row with a leading one in the leftmost column (called a *pivot*) and systematically subtracting that row from all rows below it to get zeros below in the entire column. In the second phase, we start with the rightmost pivot and use it to eliminate all the numbers above it in the same column. Let's see this in action.

1) The first step is to use the pivot in the first column to eliminate the variable x_1 in the second row. We do this by subtracting three times the first row from the second row, denoted $R_2 \leftarrow R_2 - 3R_1$,
$$\left[\begin{array}{cc|c} \mathbf{1} & 2 & 5 \\ 0 & 3 & 6 \end{array}\right].$$
2) Next, we create a pivot in the second row using $R_2 \leftarrow \frac{1}{3}R_2$:
$$\left[\begin{array}{cc|c} 1 & 2 & 5 \\ 0 & \mathbf{1} & 2 \end{array}\right].$$
3) We now start the backward phase and eliminate the second variable from the first row. We do this by subtracting two times the second row from the first row $R_1 \leftarrow R_1 - 2R_2$:
$$\left[\begin{array}{cc|c} 1 & 0 & 1 \\ 0 & 1 & 2 \end{array}\right].$$
The matrix is now in *reduced row echelon form* (RREF), which is its "simplest" form it could be in. The solutions are: $x_1 = 1$, $x_2 = 2$.

B. Systems of equations as matrix equations

We will now discuss another approach for solving the system of equations. Using the definition of the matrix-vector product, we can express this system of equations (1) as a matrix equation:

$$\begin{bmatrix} 1 & 2 \\ 3 & 9 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 21 \end{bmatrix}.$$

This matrix equation had the form $A\vec{x} = \vec{b}$, where A is a 2×2 matrix, $\vec{x}$ is the vector of unknowns, and $\vec{b}$ is a vector of constants. We can solve for $\vec{x}$ by multiplying both sides of the equation by the matrix inverse A^{-1}:

$$A^{-1}A\vec{x} = \mathbb{1}\vec{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = A^{-1}\vec{b} = \begin{bmatrix} 3 & -\frac{2}{3} \\ -1 & \frac{1}{3} \end{bmatrix}\begin{bmatrix} 5 \\ 21 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

But how did we know what the inverse matrix A^{-1} is?

IV. Computing the inverse of a matrix

In this section we'll look at several different approaches for computing the inverse of a matrix. The matrix inverse is *unique* so no matter which method we use to find the inverse, we'll always obtain the same answer.

A. Using row operations

One approach for computing the inverse is to use the Gauss–Jordan elimination procedure. Start by creating an array containing the entries of the matrix A on the left side and the identity matrix on the right side:

$$\left[\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 3 & 9 & 0 & 1 \end{array}\right].$$

Now we perform the Gauss-Jordan elimination procedure on this array.

1) The first row operation is to subtract three times the first row from the second row: $R_2 \leftarrow R_2 - 3R_1$. We obtain:
$$\left[\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 0 & 3 & -3 & 1 \end{array}\right].$$
2) The second row operation is divide the second row by 3: $R_2 \leftarrow \frac{1}{3}R_2$
$$\left[\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 0 & 1 & -1 & \frac{1}{3} \end{array}\right].$$
3) The third row operation is $R_1 \leftarrow R_1 - 2R_2$
$$\left[\begin{array}{cc|cc} 1 & 0 & 3 & -\frac{2}{3} \\ 0 & 1 & -1 & \frac{1}{3} \end{array}\right].$$

The array is now in reduced row echelon form (RREF). The inverse matrix appears on the right side of the array.

Observe that the sequence of row operations we used to solve the specific system of equations in $A\vec{x} = \vec{b}$ in the previous section are the same as the row operations we used in this section to find the inverse matrix. Indeed, in both cases the combined effect of the three row operations is to "undo" the effects of A. The right side of the 2×4 array is simply a convenient way to record this sequence of operations and thus obtain A^{-1}.

B. Using elementary matrices

Every row operation we perform on a matrix is equivalent to a left-multiplication by an *elementary matrix*. There are three types of elementary matrices in correspondence with the three types of row operations:

$$\mathcal{R}_\alpha : R_1 \leftarrow R_1 + mR_2 \quad \Leftrightarrow \quad E_\alpha = \begin{bmatrix} 1 & m \\ 0 & 1 \end{bmatrix}$$

$$\mathcal{R}_\beta : R_1 \leftrightarrow R_2 \quad \Leftrightarrow \quad E_\beta = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$\mathcal{R}_\gamma : R_1 \leftarrow mR_1 \quad \Leftrightarrow \quad E_\gamma = \begin{bmatrix} m & 0 \\ 0 & 1 \end{bmatrix}$$

Let's revisit the row operations we used to find A^{-1} in the above section representing each row operation as an elementary matrix multiplication.

1) The first row operation $R_2 \leftarrow R_2 - 3R_1$ corresponds to a multiplication by the elementary matrix E_1:
$$E_1 A = \begin{bmatrix} 1 & 0 \\ -3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 9 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix}.$$
2) The second row operation $R_2 \leftarrow \frac{1}{3}R_2$ corresponds to a matrix E_2:
$$E_2(E_1 A) = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}.$$
3) The final step, $R_1 \leftarrow R_1 - 2R_2$, corresponds to the matrix E_3:
$$E_3(E_2 E_1 A) = \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Note that $E_3 E_2 E_1 A = \mathbb{1}$, so the product $E_3 E_2 E_1$ must be equal to A^{-1}:

$$A^{-1} = E_3 E_2 E_1 = \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -3 & 1 \end{bmatrix} = \begin{bmatrix} 3 & -\frac{2}{3} \\ -1 & \frac{1}{3} \end{bmatrix}.$$

The elementary matrix approach teaches us that every invertible matrix can be decomposed as the product of elementary matrices. Since we know $A^{-1} = E_3 E_2 E_1$ then $A = (A^{-1})^{-1} = (E_3 E_2 E_1)^{-1} = E_1^{-1} E_2^{-1} E_3^{-1}$.

C. Using a computer

The last (and most practical) approach for finding the inverse of a matrix is to use a computer algebra system like the one at live.sympy.org.

```
>>> A = Matrix( [[1,2],[3,9]] )            # define A
    [1, 2]
    [3, 9]

>>> A.inv()                 # calls the inv method on A
    [ 3, -2/3]
    [-1,  1/3]
```

You can use `sympy` to "check" your answers on homework problems.

V. Other topics

We'll now discuss a number of other important topics of linear algebra.

A. Basis

Intuitively, a basis is any set of vectors that can be used as a coordinate system for a vector space. You are certainly familiar with the standard basis for the xy-plane that is made up of two orthogonal axes: the x-axis and the y-axis. A vector $\vec{v}$ can be described as a coordinate pair (v_x, v_y) with respect to these axes, or equivalently as $\vec{v} = v_x\hat{\imath} + v_y\hat{\jmath}$, where $\hat{\imath} \equiv (1, 0)$ and $\hat{\jmath} \equiv (0, 1)$ are unit vectors that point along the x-axis and y-axis respectively. However, other coordinate systems are also possible.

Definition (Basis)**.** *A basis for a n-dimensional vector space $\mathcal{S}$ is any set of n linearly independent vectors that are part of $\mathcal{S}$.*

Any set of two linearly independent vectors $\{\hat{e}_1, \hat{e}_2\}$ can serve as a basis for $\mathbb{R}^2$. We can write any vector $\vec{v} \in \mathbb{R}^2$ as a linear combination of these basis vectors $\vec{v} = v_1\hat{e}_1 + v_2\hat{e}_2$.

Note the *same* vector $\vec{v}$ corresponds to different coordinate pairs depending on the basis used: $\vec{v} = (v_x, v_y)$ in the standard basis $B_s \equiv \{\hat{\imath}, \hat{\jmath}\}$, and $\vec{v} = (v_1, v_2)$ in the basis $B_e \equiv \{\hat{e}_1, \hat{e}_2\}$. Therefore, it is important to keep in mind the basis with respect to which the coefficients are taken, and if necessary specify the basis as a subscript, e.g., $(v_x, v_y)_{B_s}$ or $(v_1, v_2)_{B_e}$.

Converting a coordinate vector from the basis B_e to the basis B_s is performed as a multiplication by a *change of basis* matrix:

$$\left[\vec{v}\right]_{B_s} = {}_{B_s}\left[\ \mathbb{1}\ \right]_{B_e} \left[\vec{v}\right]_{B_e} \quad \Leftrightarrow \quad \begin{bmatrix} v_x \\ v_y \end{bmatrix} = \begin{bmatrix} \hat{\imath}\cdot\hat{e}_1 & \hat{\imath}\cdot\hat{e}_2 \\ \hat{\jmath}\cdot\hat{e}_1 & \hat{\jmath}\cdot\hat{e}_2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}.$$

Note the change of basis matrix is actually an identity transformation. The vector $\vec{v}$ remains unchanged—it is simply expressed with respect to a new coordinate system. The change of basis from the B_s-basis to the B_e-basis is accomplished using the inverse matrix: ${}_{B_e}[\mathbb{1}]_{B_s} = ({}_{B_s}[\mathbb{1}]_{B_e})^{-1}$.

B. Matrix representations of linear transformations

Bases play an important role in the representation of linear transformations $T\colon \mathbb{R}^n \to \mathbb{R}^m$. To fully describe the matrix that corresponds to some linear transformation T, it is sufficient to know the effects of T to the n vectors of the standard basis for the input space. For a linear transformation $T\colon \mathbb{R}^2 \to \mathbb{R}^2$, the matrix representation corresponds to

$$M_T = \begin{bmatrix} | & | \\ T(\hat{\imath}) & T(\hat{\jmath}) \\ | & | \end{bmatrix} \in \mathbb{R}^{2\times 2}.$$

As a first example, consider the transformation Π_x which projects vectors onto the x-axis. For any vector $\vec{v} = (v_x, v_y)$, we have $\Pi_x(\vec{v}) = (v_x, 0)$. The matrix representation of Π_x is

$$M_{\Pi_x} = \begin{bmatrix} \Pi_x\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) & \Pi_x\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

As a second example, let's find the matrix representation of R_θ, the counterclockwise rotation by the angle θ:

$$M_{R_\theta} = \begin{bmatrix} R_\theta\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) & R_\theta\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

The first column of M_{R_θ} shows that R_θ maps the vector $\hat{\imath} \equiv 1\angle 0$ to the vector $1\angle\theta = (\cos\theta, \sin\theta)^\mathsf{T}$. The second column shows that R_θ maps the vector $\hat{\jmath} = 1\angle\frac{\pi}{2}$ to the vector $1\angle(\frac{\pi}{2} + \theta) = (-\sin\theta, \cos\theta)^\mathsf{T}$.

C. Dimension and bases for vector spaces

The *dimension* of a vector space is defined as the number of vectors in a basis for that vector space. Consider the following vector space $\mathcal{S} = \text{span}\{(1,0,0),(0,1,0),(1,1,0)\}$. Seeing that the space is described by three vectors, we might think that $\mathcal{S}$ is 3-dimensional. This is not the case, however, since the three vectors are not linearly independent so they don't form a basis for $\mathcal{S}$. Two vectors are sufficient to describe any vector in $\mathcal{S}$; we can write $\mathcal{S} = \text{span}\{(1,0,0),(0,1,0)\}$, and we see these two vectors are linearly independent so they form a basis and $\dim(\mathcal{S}) = 2$.

There is a general procedure for finding a basis for a vector space. Suppose you are given a description of a vector space in terms of m vectors $\mathcal{V} = \text{span}\{\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_m\}$ and you are asked to find a basis for $\mathcal{V}$ and the dimension of $\mathcal{V}$. To find a basis for $\mathcal{V}$, you must find a set of linearly independent vectors that span $\mathcal{V}$. We can use the Gauss–Jordan elimination procedure to accomplish this task. Write the vectors $\vec{v}_i$ as the rows of a matrix M. The vector space $\mathcal{V}$ corresponds to the row space of the matrix M. Next, use row operations to find the reduced row echelon form (RREF) of the matrix M. Since row operations do not change the row space of the matrix, the row space of reduced row echelon form of the matrix M is the same as the row space of the original set of vectors. The nonzero rows in the RREF of the matrix form a basis for vector space $\mathcal{V}$ and the numbers of nonzero rows is the dimension of $\mathcal{V}$.

D. Row space, columns space, and rank of a matrix

Recall the fundamental vector spaces for matrices that we defined in Section II-E: the column space $\mathcal{C}(A)$, the null space $\mathcal{N}(A)$, and the row space $\mathcal{R}(A)$. A standard linear algebra exam question is to give you a certain matrix A and ask you to find the dimension and a basis for each of its fundamental spaces.

In the previous section we described a procedure based on Gauss–Jordan elimination which can be used "distill" a set of linearly independent vectors which form a basis for the row space $\mathcal{R}(A)$. We will now illustrate this procedure with an example, and also show how to use the RREF of the matrix A to find bases for $\mathcal{C}(A)$ and $\mathcal{N}(A)$.

Consider the following matrix and its reduced row echelon form:

$$A = \begin{bmatrix} 1 & 3 & 3 & 3 \\ 2 & 6 & 7 & 6 \\ 3 & 9 & 9 & 10 \end{bmatrix} \qquad \text{rref}(A) = \begin{bmatrix} \mathbf{1} & 3 & 0 & 0 \\ 0 & 0 & \mathbf{1} & 0 \\ 0 & 0 & 0 & \mathbf{1} \end{bmatrix}.$$

The reduced row echelon form of the matrix A contains three pivots. The locations of the pivots will play an important role in the following steps.

The vectors $\{(1,3,0,0),(0,0,1,0),(0,0,0,1)\}$ form a basis for $\mathcal{R}(A)$.

To find a basis for the column space $\mathcal{C}(A)$ of the matrix A we need to find which of the columns of A are linearly independent. We can do this by identifying the columns which contain the leading ones in $\text{rref}(A)$. The corresponding columns in the original matrix form a basis for the column space of A. Looking at $\text{rref}(A)$ we see the first, third, and fourth columns of the matrix are linearly independent so the vectors $\{(1,2,3)^\mathsf{T}, (3,7,9)^\mathsf{T}, (3,6,10)^\mathsf{T}\}$ form a basis for $\mathcal{C}(A)$.

Now let's find a basis for the null space, $\mathcal{N}(A) \equiv \{\vec{x} \in \mathbb{R}^4 \mid A\vec{x} = \vec{0}\}$. The second column does not contain a pivot, therefore it corresponds to a *free variable*, which we will denote s. We are looking for a vector with three unknowns and one free variable $(x_1, s, x_3, x_4)^\mathsf{T}$ that obeys the conditions:

$$\begin{bmatrix} \mathbf{1} & 3 & 0 & 0 \\ 0 & 0 & \mathbf{1} & 0 \\ 0 & 0 & 0 & \mathbf{1} \end{bmatrix} \begin{bmatrix} x_1 \\ s \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad \Rightarrow \quad \begin{aligned} 1x_1 + 3s &= 0 \\ 1x_3 &= 0 \\ 1x_4 &= 0 \end{aligned}$$

Let's express the unknowns x_1, x_3, and x_4 in terms of the free variable s. We immediately see that $x_3 = 0$ and $x_4 = 0$, and we can write $x_1 = -3s$. Therefore, any vector of the form $(-3s, s, 0, 0)$, for any $s \in \mathbb{R}$, is in the null space of A. We write $\mathcal{N}(A) = \text{span}\{(-3,1,0,0)^\mathsf{T}\}$.

Observe that the $\dim(\mathcal{C}(A)) = \dim(\mathcal{R}(A)) = 3$, this is known as the *rank* of the matrix A. Also, $\dim(\mathcal{R}(A)) + \dim(\mathcal{N}(A)) = 3 + 1 = 4$, which is the dimension of the input space of the linear transformation T_A.

E. Invertible matrix theorem

There is an important distinction between matrices that are invertible and those that are not as formalized by the following theorem.

Theorem. *For an $n \times n$ matrix A, the following statements are equivalent:*

1) *A is invertible*
2) *The RREF of A is the $n \times n$ identity matrix*
3) *The rank of the matrix is n*
4) *The row space of A is $\mathbb{R}^n$*
5) *The column space of A is $\mathbb{R}^n$*
6) *A doesn't have a null space (only the zero vector $\mathcal{N}(A) = \{\vec{0}\}$)*
7) *The determinant of A is nonzero $\det(A) \neq 0$*

For a given matrix A, the above statements are either all true or all false.

An invertible matrix A corresponds to a linear transformation T_A which maps the n-dimensional input vector space $\mathbb{R}^n$ to the n-dimensional output vector space $\mathbb{R}^n$ such that there exists an inverse transformation T_A^{-1} that can faithfully undo the effects of T_A.

On the other hand, an $n \times n$ matrix B that is not invertible maps the input vector space $\mathbb{R}^n$ to a subspace $\mathcal{C}(B) \subsetneq \mathbb{R}^n$ and has a nonempty null space. Once T_B sends a vector $\vec{w} \in \mathcal{N}(B)$ to the zero vector, there is no T_B^{-1} that can undo this operation.

F. Determinants

The determinant of a matrix, denoted $\det(A)$ or $|A|$, is a special way to combine the entries of a matrix that serves to check if a matrix is invertible or not. The determinant formulas for 2×2 and 3×3 matrices are

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}, \qquad \text{and}$$

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}.$$

If the $|A| = 0$ then A is not invertible. If $|A| \neq 0$ then A is invertible.

G. Eigenvalues and eigenvectors

The set of eigenvectors of a matrix is a special set of input vectors for which the action of the matrix is described as a simple *scaling*. When a matrix is multiplied by one of its eigenvectors the output is the same eigenvector multiplied by a constant $A\vec{e}_\lambda = \lambda\vec{e}_\lambda$. The constant λ is called an *eigenvalue* of A.

To find the eigenvalues of a matrix we start from the eigenvalue equation $A\vec{e}_\lambda = \lambda\vec{e}_\lambda$, insert the identity $\mathbb{1}$, and rewrite it as a null-space problem:

$$A\vec{e}_\lambda = \lambda\mathbb{1}\vec{e}_\lambda \qquad \Rightarrow \qquad (A - \lambda\mathbb{1})\,\vec{e}_\lambda = \vec{0}.$$

This equation will have a solution whenever $|A - \lambda\mathbb{1}| = 0$. The eigenvalues of $A \in \mathbb{R}^{n \times n}$, denoted $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, are the roots of the *characteristic polynomial* $p(\lambda) = |A - \lambda\mathbb{1}|$. The *eigenvectors* associated with the eigenvalue λ_i are the vectors in the null space of the matrix $(A - \lambda_i\mathbb{1})$.

Certain matrices can be written entirely in terms of their eigenvectors and their eigenvalues. Consider the matrix Λ that has the eigenvalues of the matrix A on the diagonal, and the matrix Q constructed from the eigenvectors of A as columns:

$$\Lambda = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & 0 \\ 0 & 0 & \lambda_n \end{bmatrix}, \quad Q = \begin{bmatrix} | & & | \\ \vec{e}_{\lambda_1} & \cdots & \vec{e}_{\lambda_n} \\ | & & | \end{bmatrix}, \quad \text{then} \quad A = Q\Lambda Q^{-1}.$$

Matrices that can be written this way are called diagonalizable.

The decomposition of a matrix into its eigenvalues and eigenvectors gives valuable insights into the properties of the matrix. Google's original `PageRank` algorithm for ranking webpages by "importance" can be formalized as an eigenvector calculation on the matrix of web hyperlinks.

VI. Textbook plug

Math 54 Cheat Sheet

Vector spaces

Subspace: If $\mathbf{u}$ and $\mathbf{v}$ are in W, then $\mathbf{u}+\mathbf{v}$ are in W, and $c\mathbf{u}$ is in W
Nul(A): Solutions of $A\mathbf{x}=\mathbf{0}$. Row-reduce A.
Row(A): Space spanned by the rows of A: Row-reduce A and choose the rows that contain the pivots.
Col(A): Space spanned by columns of A: Row-reduce A and choose the **columns of** A that contain the pivots
Rank(A): = $Dim(Col(A))$ = number of pivots
Rank-Nullity theorem: $Rank(A)+dim(Nul(A))=n$, where A is $m\times n$
Linear transformation: $T(\mathbf{u}+\mathbf{v})=T(\mathbf{u})+T(\mathbf{v}), T(c\mathbf{u})=cT(\mathbf{u})$, where c is a number.
T is one-to-one if $T(\mathbf{u})=\mathbf{0}\Rightarrow\mathbf{u}=\mathbf{0}$
T is onto if $Col(T)=\mathbb{R}^m$.
Linearly independence: $a_1\mathbf{v_1}+a_2\mathbf{v_2}+\cdots+a_n\mathbf{v_n}=\mathbf{0}\Rightarrow a_1=a_2=\cdots=a_n=0$.
To show lin. ind, form the matrix of the vectors, and show that $Nul(A)=\{\mathbf{0}\}$
Linear dependence: $a_1\mathbf{v_1}+a_2\mathbf{v_2}+\cdots+a_n\mathbf{v_n}=\mathbf{0}$ for $a_1, a_2, \cdots, a_n$, not all zero.
Span: Set of linear combinations of $\mathbf{v_1}, \cdots \mathbf{v_n}$
Basis $\mathcal{B}$ for V: A linearly independent set such that $Span(\mathcal{B})=V$
To show sthg is a basis, show it is linearly independent and spans.
To find a basis from a collection of vectors, form the matrix A of the vectors, and find $Col(A)$.
To find a basis for a vector space, take any element of that v.s. and express it as a linear combination of 'simpler' vectors. Then show those vectors form a basis.
Dimension: Number of elements in a basis.
To find dim, find a basis and find num. elts.
Theorem: If V has a basis of vectors, then every basis of V must have n vectors.
Basis theorem: If V is an $n-$dim v.s., then any lin. ind. set with n elements is a basis, and any set of n elts. which spans V is a basis. Matrix of a lin. transf T with respect to bases $\mathcal{B}$ and $\mathcal{C}$: For every vector $\mathbf{v}$ in $\mathcal{B}$, evaluate $T(\mathbf{v})$, and express $T(\mathbf{v})$ as a linear combination of vectors in $\mathcal{C}$. Put the **coefficients** in a column vector, and then form the matrix of the column vectors you found!
Coordinates: To find $[\mathbf{x}]_{\mathcal{B}}$, express $\mathbf{x}$ in terms of the vectors in $\mathcal{B}$.
$\mathbf{x}=P_{\mathcal{B}}[\mathbf{x}]_{\mathcal{B}}$, where $P_{\mathcal{B}}$ is the matrix whole columns are the vectors in $\mathcal{B}$.
Invertible matrix theorem: If A is invertible, then: A is row-equivalent to I, A has n pivots, $T(\mathbf{x})=A\mathbf{x}$ is one-to-one and onto, $A\mathbf{x}=\mathbf{b}$ has a unique solution for every $\mathbf{b}$, A^T is invertible, $det(A)\neq 0$, the columns of A form a basis for $\mathbb{R}^n$, $Nul(A)=\{\mathbf{0}\}$, $Rank(A)=n$
$\begin{bmatrix}a & b\\ c & d\end{bmatrix}^{-1}=\frac{1}{ad-bc}\begin{bmatrix}d & -b\\ -c & a\end{bmatrix}$
$[A\mid I]\rightarrow\left[I\mid A^{-1}\right]$
Change of basis: $[\mathbf{x}]_{\mathcal{C}}=P_{\mathcal{C}\leftarrow\mathcal{B}}[\mathbf{x}]_{\mathcal{B}}$ (think of $\mathcal{C}$ as the new, cool basis)
$[\mathcal{C}\mid\mathcal{B}]\rightarrow[I\mid P_{\mathcal{C}\leftarrow\mathcal{B}}]$
$P_{\mathcal{C}\leftarrow\mathcal{B}}$ is the matrix whose columns are $[\mathbf{b}]_{\mathcal{C}}$, where $\mathbf{b}$ is in $\mathcal{B}$

Diagonalization

Diagonalizability: A is **diagonalizable** if $A=PDP^{-1}$ for some diagonal D and invertible P.
A and B are similar if $A=PBP^{-1}$ for P invertible
Theorem: A is diagonalizable $\Leftrightarrow$ A has n linearly independent **eigenvectors**
Theorem: **IF** A has n distinct eigenvalues, **THEN** A is diagonalizable, but the opposite is not always true!!!!
Notes: A can be diagonalizable even if it's not invertible (Ex: $A=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}$). Not all matrices are diagonalizable (Ex: $\begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix}$)
Consequence: $A=PDP^{-1}\Rightarrow A^n=PD^nP^{-1}$
How to diagonalize: To find the eigenvalues, calculate $det(A-\lambda I)$, and find the roots of that.
To find the eigenvectors, for each λ find a basis for $Nul(A-\lambda I)$, which you do by row-reducing
Rational roots theorem: If $p(\lambda)=0$ has a rational root $r=\frac{a}{b}$, then a divides the constant term of p, and b divides the leading coefficient.
Use this to guess zeros of p. Once you have a zero that works, use long division! Then $A=PDP^{-1}$, where D= diagonal matrix of eigenvalues, P = matrix of eigenvectors
Complex eigenvalues If $\lambda=a+bi$, and $\mathbf{v}$ is an eigenvector, then $A=PCP^{-1}$, where
$P=[Re(\mathbf{v})\quad Im(\mathbf{v})], C=\begin{bmatrix}a & b\\ -b & a\end{bmatrix}$
C is a scaling of $\sqrt{det(A)}$ followed by a rotation by θ, where:
$\frac{1}{\sqrt{det(A)}}C=\begin{bmatrix}\cos(\theta) & \sin(\theta)\\ -\sin(\theta) & \cos(\theta)\end{bmatrix}$

Orthogonality

$\mathbf{u}, \mathbf{v}$ orthogonal if $\mathbf{u}\cdot\mathbf{v}=0$.
$\|\mathbf{u}\|=\sqrt{\mathbf{u}\cdot\mathbf{u}}$
$\{\mathbf{u_1}\cdots\mathbf{u_n}\}$ is orthogonal if $\mathbf{u_i}\cdot\mathbf{u_j}=0$ if $i\neq j$, orthonormal if $\mathbf{u_i}\cdot\mathbf{u_i}=1$
$W^{\perp}$: Set of $\mathbf{v}$ which are orthogonal to every $\mathbf{w}$ in W.
If $\{\mathbf{u_1}\cdots\mathbf{u_n}\}$ is an orthogonal basis, then:
$\mathbf{y}=c_1\mathbf{u_1}+\cdots c_n\mathbf{u_n}\Rightarrow c_j=\frac{\mathbf{y}\cdot\mathbf{u_j}}{\mathbf{u_j}\cdot\mathbf{u_j}}$
Orthogonal matrix Q has ortho**normal** columns! Consequence: $Q^TQ=I$, $QQ^T=$ Orthogonal projection on $Col(Q)$.
$\|Q\mathbf{x}\|=\|\mathbf{x}\|$
$(Q\mathbf{x})\cdot(Q\mathbf{y})=\mathbf{x}\cdot\mathbf{y}$
Orthogonal projection: If $\{\mathbf{u_1}\cdots\mathbf{u_k}\}$ is a basis for W, then orthogonal projection of $\mathbf{y}$ on W is:
$\hat{\mathbf{y}}=\left(\frac{\mathbf{y}\cdot\mathbf{u_1}}{\mathbf{u_1}\mathbf{u_1}}\right)\mathbf{u_1}+\cdots+\left(\frac{\mathbf{y}\cdot\mathbf{u_1}}{\mathbf{u_k}\mathbf{u_k}}\right)\mathbf{u_k}$
$\mathbf{y}-\hat{\mathbf{y}}$ is orthogonal to $\hat{\mathbf{y}}$, shortest distance btw $\mathbf{y}$ and W is $\|\mathbf{y}-\hat{\mathbf{y}}\|$
Gram-Schmidt: Start with $\mathcal{B}=\{\mathbf{u_1}, \cdots \mathbf{u_n}\}$. Let:
$\mathbf{v_1}=\mathbf{u_1}$
$\mathbf{v_2}=\mathbf{u_2}-\left(\frac{\mathbf{u_2}\cdot\mathbf{v_1}}{\mathbf{v_1}\cdot\mathbf{v_1}}\right)\mathbf{v_1}$
$\mathbf{v_3}=\mathbf{u_3}-\left(\frac{\mathbf{u_3}\cdot\mathbf{v_1}}{\mathbf{v_1}\cdot\mathbf{v_1}}\right)\mathbf{v_1}-\left(\frac{\mathbf{u_3}\cdot\mathbf{v_2}}{\mathbf{v_2}\cdot\mathbf{v_2}}\right)\mathbf{v_2}$
Then $\{\mathbf{v_1}\cdots\mathbf{v_n}\}$ is an orthogonal basis for $Span(\mathcal{B})$, and if $\mathbf{w_i}=\frac{\mathbf{v_i}}{\|\mathbf{v_i}\|}$, then $\{\mathbf{w_1}\cdots\mathbf{w_n}\}$ is an orthonormal basis for $Span(\mathcal{B})$.
QR-factorization: To find Q, apply G-S to columns of A. Then $R=Q^TA$
Least-squares: To solve $A\mathbf{x}=\mathbf{b}$ in the least squares-way, solve $A^TA\mathbf{x}=A^T\mathbf{b}$.
Least squares solution makes $\|A\mathbf{x}-\mathbf{b}\|$ smallest.
$\hat{\mathbf{x}}=R^{-1}Q^T\mathbf{b}$, where $A=QR$.
Inner product spaces $f\cdot g=\int_a^b f(t)g(t)dt$. G-S applies with this inner product as well.
Cauchy-Schwarz: $|\mathbf{u}\cdot\mathbf{v}|\leq\|\mathbf{u}\|\,\|\mathbf{v}\|$
Triangle inequality: $\|\mathbf{u}+\mathbf{v}\|\leq\|\mathbf{u}\|+\|\mathbf{v}\|$

Symmetric matrices ($A=A^T$)

Has n real eigenvalues, always diagonalizable, orthogonally diagonalizable ($A=PDP^T$, P is an orthogonal matrix, equivalent to symmetry!).
Theorem: If A is symmetric, then any two eigenvectors from different eigenspaces are orthogonal.
How to orthogonally diagonalize: First diagonalize, then apply G-S on each eigenspace and normalize. Then P = matrix of (orthonormal) eigenvectors, D = matrix of eigenvalues.
Quadratic forms: To find the matrix, put the x_i^2-coefficients on the diagonal, and evenly distribute the other terms. For example, if the x_1x_2-term is 6, then the $(1,2)$th and $(2,1)$th entry of A is 3.
Then orthogonally diagonalize $A=PDP^T$.
Then let $\mathbf{y}=P^T\mathbf{x}$, then the quadratic form becomes $\lambda_1y_1^2+\cdots+\lambda_ny_n^2$, where λ_i are the eigenvalues.
Spectral decomposition: $\lambda_1\mathbf{u_1}\mathbf{u_1}^T+\lambda_2\mathbf{u_2}\mathbf{u_2}^T+\cdots+\lambda_n\mathbf{u_n}\mathbf{u_n}^T$

Second-order and Higher-order differential equations

Homogeneous solutions: Auxiliary equation: Replace equation by polynomial, so y''' becomes r^3 etc. Then find the zeros (use the rational roots theorem and long division, see the 'Diagonalization-section). 'Simple zeros' give you e^{rt}, Repeated zeros (multiplicity m) give you $Ae^{rt}+Bte^{rt}+\cdots Zt^{m-1}e^{rt}$, Complex zeros $r=a+bi$ give you $Ae^{at}\cos(bt)+Be^{at}\sin(bt)$.
Undetermined coefficients: $y(t)=y_0(t)+y_p(t)$, where y_0 solves the hom. eqn. (equation $=0$), and y_p is a particular solution. To find y_p:
If the inhom. term is Ct^me^{rt}, then: $y_p=t^s(A_mt^m\cdots+A_1t+1)e^{rt}$, where if r is a root of aux with multiplicity m, then $s=m$, and if r is not a root, then $s=0$.
If the inhom term is $Ct^me^{at}\sin(\beta t)$, then:
$y_p=t^s(A_mt^m\cdots+A_1t+1)e^{at}\cos(\beta t)+t^s(B_mt^m\cdots+B_1t+1)e^{rt}\sin(\beta t)$, where $s=m$, if $a+bi$ is also a root of aux with multiplicity m ($s=0$ if not). $\cos$ **always goes with** $\sin$ **and vice-versa**, also, you have to look at $a+bi$ as one entity.
Variation of parameters: **First, make sure the leading coefficient (usually the coeff. of y'') is** $=1$.. Then $y=y_0+y_p$ as above. Now suppose $y_p(t)=v_1(t)y_1(t)+v_2(t)y_2(t)$, where y_1 and y_2 are your hom. solutions. Then $\begin{bmatrix}y_1 & y_2\\ y_1' & y_2'\end{bmatrix}\begin{bmatrix}v_1'\\ v_2'\end{bmatrix}=\begin{bmatrix}0\\ f(t)\end{bmatrix}$. Invert the matrix and solve for v_1' and v_2', and integrate to get v_1 and v_2, and finally use: $y_p(t)=v_1(t)y_1(t)+v_2(t)y_2(t)$.
Useful formulas: $\begin{bmatrix}a & b\\ c & d\end{bmatrix}^{-1}=\frac{1}{ad-bc}\begin{bmatrix}d & -b\\ -c & a\end{bmatrix}$
$\int\sec(t)=\ln|\sec(t)+\tan(t)|, \int\tan(t)=\ln|\sec(t)|, \int\tan^2(t)=\tan(x)-x$,
$\int\ln(t)=t\ln(t)-t$
Linear independence: f, g, h are linearly independent if
$af(t)+bg(t)+ch(t)=0\Rightarrow a=b=c=0$. To show linear dependence, do it directly. To show linear independence, form the Wronskian: $\widetilde{W}(t)=\begin{bmatrix}f(t) & g(t)\\ f'(t) & g'(t)\end{bmatrix}$ (for 2 functions),
$\widetilde{W}(t)=\begin{bmatrix}f(t) & g(t) & h(t)\\ f'(t) & g'(t) & h'(t)\\ f''(t) & g''(t) & h''(t)\end{bmatrix}$ (for 3 functions). Then pick a point t_0 where $det(\widetilde{W}(t_0))$ is easy to evaluate. If $det\neq 0$, then f, g, h are linearly independent! Try to look for simplifications before you differentiate.
Fundamental solution set: If f, g, h are solutions **and** linearly independent.
Largest interval of existence: First make sure the leading coefficient equals to 1. Then look at the domain of each term. For each domain, consider the part of the interval which contains the initial condition. Finally, intersect the intervals and change any brackets to parentheses. Harmonic oscillator: $my''+by'+ky=0$ ($m=$ inertia, $b=$ damping, $k=$ stiffness)

Systems of differential equations

To solve $\mathbf{x}'=A\mathbf{x}$: $\mathbf{x}(t)=Ae^{\lambda_1t}\mathbf{v_1}+Be^{\lambda_2t}\mathbf{v_2}+e^{\lambda_3t}\mathbf{v_3}$ (λ_i are your eigenvalues, $\mathbf{v_i}$ are your eigenvectors)
Fundamental matrix: Matrix whose columns are the solutions, without the constants (the columns are solutions and linearly independent)
Complex eigenvalues If $\lambda=\alpha+i\beta$, and $\mathbf{v}=\mathbf{a}+i\mathbf{b}$. Then:
$\mathbf{x}(t)=A\left(e^{\alpha t}\cos(\beta t)\mathbf{a}-e^{\alpha t}\sin(\beta t)\mathbf{b}\right)+B\left(e^{\alpha t}\sin(\beta t)\mathbf{a}+e^{\alpha t}\cos(\beta t)\mathbf{b}\right)$
Notes: You only need to consider one complex eigenvalue. For real eigenvalues, use the formula above. Also, $\frac{1}{a+bi}=\frac{a-bi}{a^2+b^2}$
Generalized eigenvectors If you only find one eigenvector $\mathbf{v}$ (even though there are supposed to be 2), then solve the following equation for $\mathbf{u}$: $(A-\lambda I)(\mathbf{u})=\mathbf{v}$ (one solution is enough).
Then: $\mathbf{x}(t)=Ae^{\lambda t}\mathbf{v}+B\left(te^{\lambda t}\mathbf{v}+e^{\lambda t}\mathbf{u}\right)$
Undetermined coefficients First hom. solution. Then for $\mathbf{x_p}$, just like regular undetermined coefficients, except that instead of guessing $\mathbf{x_p}(t)=ae^t+b\cos(t)$, you guess $\mathbf{a}e^t+\mathbf{b}\cos(t)$, where $\mathbf{a}=\begin{bmatrix}a_1\\ a_2\end{bmatrix}$ is a vector. Then plug into $\mathbf{x}'=A\mathbf{x}+\mathbf{f}$ and solve for $\mathbf{a}$ etc.
Variation of parameters First hom. solution $\mathbf{x_h}(t)=A\mathbf{x_1}(t)+B\mathbf{x_2}(t)$. Then sps
$\mathbf{x_p}(t)=v_1(t)\mathbf{x_1}(t)+v_2(t)\mathbf{x_2}(t)$, then solve $\widetilde{W}(t)\begin{bmatrix}v_1'\\ v_2'\end{bmatrix}=\mathbf{f}$, where
$\widetilde{W}(t)=[\mathbf{x_1}(t)\mid\quad\mathbf{x_2}(t)]$. Multiply both sides by $\left(\widetilde{W}(t)\right)^{-1}$, integrate and solve for $v_1(t)$, $v_2(t)$, and plug back into $\mathbf{x_p}$. Finally, $\mathbf{x}=\mathbf{x_h}+\mathbf{x_p}$
Matrix exponential $e^{At}=\sum_{n=0}^{\infty}\frac{A^nt^n}{n!}$. To calculate e^{At}, either diagonalize:
$A=PDP^{-1}\Rightarrow e^{At}=Pe^{Dt}P^{-1}$, where e^{Dt} is a diagonal matrix with diag. entries e^{λ_it}. Or if A only has one eigenvalue λ with multiplicity m, use $e^{At}=e^{\lambda t}\sum_{n=0}^{m-1}\frac{(A-\lambda I)^nt^n}{n!}$. Solution of $\mathbf{x}'=A\mathbf{x}$ is then $\mathbf{x}(t)=e^{At}\mathbf{c}$, where $\mathbf{c}$ is a constant vector.

Coupled mass-spring system

Case $N=2$
Equation: $\mathbf{x}''=A\mathbf{x}, A=\begin{bmatrix}-2 & 1\\ 1 & -2\end{bmatrix}$
Proper frequencies: Eigenvalues of A are: $\lambda=-1, -3$, then proper frequencies $\boxed{\pm i, \pm\sqrt{3}i}$ ($\pm$ square roots of eigenvalues)
Proper modes: $\mathbf{v_1}=\begin{bmatrix}\sin\left(\frac{\pi}{3}\right)\\ \sin\left(2\frac{\pi}{3}\right)\end{bmatrix}=\begin{bmatrix}\frac{\sqrt{3}}{2}\\ \frac{\sqrt{3}}{2}\end{bmatrix}, \mathbf{v_2}=\begin{bmatrix}\sin\left(2\frac{\pi}{3}\right)\\ \sin\left(4\frac{\pi}{3}\right)\end{bmatrix}=\begin{bmatrix}\frac{\sqrt{3}}{2}\\ -\frac{\sqrt{3}}{2}\end{bmatrix}$

Case $N=3$
Equation: $\mathbf{x}''=A\mathbf{x}, A=\begin{bmatrix}-2 & 1 & 0\\ 1 & -2 & 1\\ 0 & 1 & -2\end{bmatrix}$
Proper frequencies: Eigenvalues of A: $\lambda=-2, -2-\sqrt{2}, -2+\sqrt{2}$, then proper frequencies
$\boxed{\pm\sqrt{2}i, \pm\left(\sqrt{2+\sqrt{2}}\right)i, \pm\left(\sqrt{2-\sqrt{2}}\right)i}$
Proper modes: $\mathbf{v_1}=\begin{bmatrix}\sin\left(\frac{\pi}{4}\right)\\ \sin\left(2\frac{\pi}{4}\right)\\ \sin\left(3\frac{\pi}{4}\right)\end{bmatrix}=\begin{bmatrix}\frac{\sqrt{2}}{2}\\ 1\\ \frac{\sqrt{2}}{2}\end{bmatrix}, \mathbf{v_2}=\begin{bmatrix}\sin\left(2\frac{\pi}{4}\right)\\ \sin\left(4\frac{\pi}{4}\right)\\ \sin\left(6\frac{\pi}{4}\right)\end{bmatrix}=\begin{bmatrix}1\\ 0\\ -1\end{bmatrix}, \mathbf{v_3}=\begin{bmatrix}\sin\left(3\frac{\pi}{4}\right)\\ \sin\left(6\frac{\pi}{4}\right)\\ \sin\left(9\frac{\pi}{4}\right)\end{bmatrix}=\begin{bmatrix}\frac{\sqrt{2}}{2}\\ -1\\ \frac{\sqrt{2}}{2}\end{bmatrix}$
General case (just in case!)
Equation: $\mathbf{x}''=A\mathbf{x}, A=\begin{bmatrix}-2 & 1 & 0 & \cdots & 0 & \\ 1 & -2 & 1 & 0 & \cdots & 0\\ 0 & 1 & -2 & 1 & 0 & \cdots\\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\ 0 & 0 & \cdots & 1 & -2 & 1\\ 0 & 0 & 0 & 0 & 1 & -2\end{bmatrix}$
Proper frequencies: $\boxed{\pm 2i\sin\left(\frac{k\pi}{2(N+1)}\right), k=1, 2, \cdots N}$
Proper modes: $\boxed{\mathbf{v_k}=\begin{bmatrix}\sin\left(\frac{k\pi}{N+1}\right)\\ \sin\left(\frac{2k\pi}{N+1}\right)\\ \cdot\\ \cdot\\ \cdot\\ \sin\left(\frac{Nk\pi}{N+1}\right)\end{bmatrix}}$

Partial differential equations

Full Fourier series: f defined on $(-T, T)$:
$f(x)\sim\sum_{m=0}^{\infty}\left(a_m\cos\left(\frac{\pi mx}{T}\right)+b_m\sin\left(\frac{\pi mx}{T}\right)\right)$, where:
$a_0=\frac{1}{2T}\int_{-T}^{T}f(x)dx$
$a_m=\frac{1}{T}\int_{-T}^{T}f(x)\cos\left(\frac{\pi mx}{T}\right)$
$b_0=0$
$b_m=\frac{1}{T}\int_{-T}^{T}f(x)\sin\left(\frac{\pi mx}{T}\right)$
Cosine series: f defined on $(0, T)$: $f(x)\sim\sum_{m=0}^{\infty}a_m\cos\left(\frac{\pi mx}{T}\right)$, where:
$a_0=\frac{2}{2T}\int_0^T f(x)dx$ (not a typo)
$a_m=\frac{2}{T}\int_0^T f(x)\cos\left(\frac{\pi mx}{T}\right)$

Sine series: f defined on $(0,\,T)$: $f(x) \sim \sum_{m=0}^{\infty} b_m \sin\left(\frac{\pi m x}{T}\right)$, where:

$b_0 = 0$

$b_m = \frac{2}{T}\int_0^T f(x)\sin\left(\frac{\pi m x}{T}\right)$

Tabular integration: (IBP: $\int f'g = fg - \int fg'$) To integrate $\int f(t)g(t)dt$ where f is a polynomial, make a table whose first row is $f(t)$ and $g(t)$. Then differentiate f as many times until you get 0, and antidifferentiate as many times until it aligns with the 0 for f. Then multiply the diagonal terms and do $+$ first term $-$ second term etc.

Orthogonality formulas: $\int_{-T}^{T}\cos\left(\frac{\pi m x}{T}\right)\sin\left(\frac{\pi n x}{T}\right)dx = 0$

$\int_{-T}^{T}\cos\left(\frac{\pi m x}{T}\right)\cos\left(\frac{\pi n x}{T}\right)dx = 0$ if $m \neq n$

$\int_{-T}^{T}\sin\left(\frac{\pi m x}{T}\right)\sin\left(\frac{\pi n x}{T}\right)dx = 0$ if $m \neq n$

Convergence: Fourier series $\mathcal{F}$ goes to $f(x)$ is f is continuous at x, and if f has a jump at x, $\mathcal{F}$ goes to the average of the jumps. Finally, at the endpoints, $\mathcal{F}$ goes to average of the left/right endpoints.

Heat/Wave equations:

Step 1: Suppose $u(x,\,t) = X(x)T(t)$, plug this into PDE, and group X-terms and T-terms. Then $\frac{X''(x)}{X(x)} = \lambda$, so $X'' = \lambda X$. Then find a differential equation for T. **Note:** If you have an α-term, put it with T.

Step 2: Deal with $X'' = \lambda X$. Use boundary conditions to find $X(0)$ etc. (if you have $\frac{\partial u}{\partial x}$, you might have $X'(0)$ instead of $X(0)$).

Step 3: Case 1: $\lambda = \omega^2$, then $X(x) = Ae^{\omega x} + Be^{-\omega x}$, then find $\omega = 0$, contradiction. Case 2: $\lambda = 0$, then $X(x) = Ax + B$, then eihter find $X(x) = 0$ (contradiction), or find $X(x) = A$. Case 3: $\lambda = -\omega^2$, then $X(x) = A\cos(\omega x) + B\sin(\omega x)$. Then solve for ω, usually $\omega = \frac{\pi m}{T}$. Also, if case 2 works, should find $\cos$, if case 2 doesn't work, should find $\sin$.
Finally, $\lambda = -\omega^2$, and $X(x) =$ whatever you found in 2) w/o the constant.

Step 4: Solve for $T(t)$ with the λ you found. Remember that for the heat equation:
$T' = \lambda T \Rightarrow T(t) = \widetilde{A_m}e^{\lambda t}$. And for the wave equation:
$T'' = \lambda T \Rightarrow T(t) = \widetilde{A_m}\cos(\omega t) + \widetilde{B_m}\sin(\omega t)$.

Step 5: Then $u(x,\,t) = \sum_{m=0}^{\infty} T(t)X(x)$ (if case 2 works), $u(x,\,t) = \sum_{m=1}^{\infty} T(t)X(x)$ (if case 2 doesn't work!)

Step 6: Use $u(x,\,0)$, and plug in $t = 0$. Then use Fourier cosine or sine series or just 'compare', i.e. if $u(x,\,0) = 4\sin(2\pi x) + 3\sin(3\pi x)$, then $\widetilde{A_2} = 4$, $\widetilde{A_3} = 3$, and $\widetilde{A_m} = 0$ if $m \neq 2,\,3$.

Step 7: (only for wave equation): Use $\frac{\partial u}{\partial t}u(x,\,0)$: Differentiate Step 5 with respect to t and set $t = 0$. Then use Fourier cosine or series or 'compare'

Nonhomogeneous heat equation: $\begin{cases} \frac{\partial u}{\partial t} = \beta\frac{\partial^2 u}{\partial x^2} + P(x) \\ u(0,\,t) = U_1, \qquad u(L,\,t) = U2 \\ u(x,\,0) = f(x) \end{cases}$

Then $u(x,\,t) = v(x) + w(x,\,t)$, where:
$v(x) = \left[U_2 - U_1 + \int_0^L \int_0^z \frac{1}{\beta}P(s)dsdz\right]\frac{x}{L} + U_1 - \int_0^x \int_0^z \frac{1}{\beta}P(s)dsdz$ and $w(x,\,t)$ solves the hom. eqn: $\begin{cases} \frac{\partial w}{\partial t} = \beta\frac{\partial^2 w}{\partial x^2} \\ w(0,\,t) = 0, \qquad w(L,\,t) = 0 \\ u(x,\,0) = f(x) - v(x) \end{cases}$

D'Alembert's formula: ONLY works for wave equation and $-\infty < x < \infty$:
$u(x,\,t) = \frac{1}{2}\left(f(x+\alpha t) + f(x-\alpha t)\right) + \frac{1}{2\alpha}\int_{x-\alpha t}^{x+\alpha t} g(s)ds$, where
$u_{tt} = \alpha^2 u_{xx}$, $u(x,\,0) = f(x)$, $\frac{\partial u}{\partial t}u(x,\,0) = g(x)$. The integral just means 'antidifferentiate and plug in'.

Laplace equation:

Same as for Heat/Wave, but $T(t)$ becomes $Y(y)$, and we get $Y''(y) = -\lambda Y(y)$. Also, instead of writing $Y(y) = \widetilde{A_m}e^{\omega y} + \widetilde{B_m}e^{-\omega y}$, write $Y(y) = \widetilde{A_m}\cosh(\omega y) + \widetilde{B_m}\sinh(\omega y)$.
Remember $\cosh(0) = 1, \sinh(0) = 0$

G00TE1204: Convex Optimization and Its Applications in Signal Processing

Handout A: Linear Algebra Cheat Sheet

Instructor: Anthony Man–Cho So Updated: May 10, 2015

The purpose of this handout is to give a brief review of some of the basic concepts and results in linear algebra. If you are not familiar with the material and/or would like to do some further reading, you may consult, e.g., the books [1, 2, 3].

1 Basic Notations, Definitions and Results

1.1 Vectors and Matrices

We denote the set of real numbers (also referred to as **scalars**) by $\mathbb{R}$. For positive integers $m, n \geq 1$, we use $\mathbb{R}^{m\times n}$ to denote the set of $m \times n$ arrays whose components are from $\mathbb{R}$. In other words, $\mathbb{R}^{m\times n}$ is the set of n–dimensional real matrices, and an element $A \in \mathbb{R}^{m\times n}$ can be written as

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \tag{1}$$

where $a_{ij} \in \mathbb{R}$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. A **row vector** is a matrix with $m = 1$, and a **column vector** is a matrix with $n = 1$. The word *vector* will always mean a column vector unless otherwise stated. The set of all n–dimensional real vectors is denoted by $\mathbb{R}^n$, and an element $x \in \mathbb{R}^n$ can be written as $x = (x_1, \ldots, x_n)$. Note that we still view $x = (x_1, \ldots, x_n)$ as a column vector, even though typographically it does not appear so. The reason for such a notation is simply to save space. Now, given an $m \times n$ matrix A of the form (1), its **transpose** A^T is defined as the following $n \times m$ matrix:

$$A^T = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix}.$$

An $m \times m$ real matrix A is said to be **symmetric** if $A = A^T$. The set of $m \times m$ real symmetric matrices is denoted by $\mathcal{S}^m$.

We use $x \geq \mathbf{0}$ to indicate that all the components of x are non–negative, and $x \geq y$ to mean that $x - y \geq \mathbf{0}$. The notations $x > \mathbf{0}$, $x \leq \mathbf{0}$, $x < \mathbf{0}$, $x > y$, $x \leq y$, and $x < y$ are to be interpreted accordingly.

We say that a finite collection $\mathcal{C} = \{x^1, x^2, \ldots, x^m\}$ of vectors in $\mathbb{R}^n$ is

- **linearly dependent** if there exist scalars $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$, not all of them zero, such that $\sum_{i=1}^m \alpha_i x^i = \mathbf{0}$;
- **affinely dependent** if the collection $\mathcal{C}' = \{x^2 - x^1, x^3 - x^1, \ldots, x^m - x^1\}$ is linearly dependent.

The collection $\mathcal{C}$ (resp. $\mathcal{C}'$) is said to be **linearly independent** (resp. **affinely independent**) if it is not linearly dependent (resp. affinely dependent).

1.2 Inner Product and Vector Norms

Given two vectors $x, y \in \mathbb{R}^n$, their **inner product** is defined as

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

We say that x and y are **orthogonal** if $x^T y = 0$. The **Euclidean norm** of $x \in \mathbb{R}^n$ is defined as

$$\|x\|_2 \equiv \sqrt{x^T x} = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2}.$$

A fundamental inequality that relates the inner product of two vectors and their respective Euclidean norms is the **Cauchy–Schwarz inequality**:

$$\left| x^T y \right| \leq \|x\|_2 \cdot \|y\|_2.$$

Equality holds iff the vectors x and y are linearly dependent; i.e., $x = \alpha y$ for some $\alpha \in \mathbb{R}$.

Note that the Euclidean norm is not the only norm one can define on $\mathbb{R}^n$. In general, a function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ is called a **vector norm** on $\mathbb{R}^n$ if for all $x, y \in \mathbb{R}^n$, we have

(a) **(Non–Negativity)** $\|x\| \geq 0$;

(b) **(Positivity)** $\|x\| = 0$ iff $x = \mathbf{0}$;

(c) **(Homogeneity)** $\|\alpha x\| = |\alpha| \cdot \|x\|$ for all $\alpha \in \mathbb{R}$;

(d) **(Triangle Inequality)** $\|x + y\| \leq \|x\| + \|y\|$.

For instance, for $p \geq 1$, the ℓ_p**–norm** on $\mathbb{R}^n$, which is given by

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p},$$

is a vector norm on $\mathbb{R}^n$. It is well known that

$$\|x\|_\infty = \lim_{p \to \infty} \|x\|_p = \max_{1 \leq i \leq n} |x_i|.$$

1.3 Matrix Norms

We say that a function $\|\cdot\| : \mathbb{R}^{n \times n} \to \mathbb{R}$ is a **matrix norm** on the set of $n \times n$ matrices if for any $A, B \in \mathbb{R}^{n \times n}$, we have

(a) **(Non–Negativity)** $\|A\| \geq 0$;

(b) **(Positivity)** $\|A\| = 0$ iff $A = \mathbf{0}$;

(c) **(Homogeneity)** $\|\alpha A\| = |\alpha| \cdot \|A\|$ for all $\alpha \in \mathbb{R}$;

(d) **(Triangle Inequality)** $\|A + B\| \leq \|A\| + \|B\|$;

(e) **(Submultiplicativity)** $\|AB\| \leq \|A\| \cdot \|B\|$.

As an example, let $\|\cdot\|_v : \mathbb{R}^n \to \mathbb{R}$ be a vector norm on $\mathbb{R}^n$. Define the function $\|\cdot\| : \mathbb{R}^{n\times n} \to \mathbb{R}$ via

$$\|A\| = \max_{x\in\mathbb{R}^n : \|x\|_v = 1} \|Ax\|_v.$$

Then, it is straightforward to verify that $\|\cdot\|$ is a matrix norm on the set of $n \times n$ matrices.

1.4 Linear Subspaces and Bases

A non–empty subset S of $\mathbb{R}^n$ is called a **(linear) subspace** of $\mathbb{R}^n$ if $\alpha x + \beta y \in S$ whenever $x, y \in S$ and $\alpha, \beta \in \mathbb{R}$. Clearly, we have $\mathbf{0} \in S$ for any subspace S of $\mathbb{R}^n$.

The **span** (or **linear hull**) of a finite collection $\mathcal{C} = \{x^1, \ldots, x^m\}$ of vectors in $\mathbb{R}^n$ is defined as

$$\text{span}(\mathcal{C}) \equiv \left\{ \sum_{i=1}^{m} \alpha_i x^i : \alpha_1, \ldots, \alpha_m \in \mathbb{R} \right\}.$$

In particular, every vector $y \in \text{span}(\mathcal{C})$ is a **linear combination** of the vectors in $\mathcal{C}$. It is easy to verify that $\text{span}(\mathcal{C})$ is a subspace of $\mathbb{R}^n$.

We can extend the above definition to an arbitrary (i.e., not necessarily finite) collection $\mathcal{C}$ of vectors in $\mathbb{R}^n$. Specifically, we define $\text{span}(\mathcal{C})$ as the set of all *finite* linear combinations of the vectors in $\mathcal{C}$. Equivalently, we can define $\text{span}(\mathcal{C})$ as the intersection of all subspaces containing $\mathcal{C}$. Note that when $\mathcal{C}$ is finite, this definition coincides with the one given above.

Given a subspace S of $\mathbb{R}^n$ with $S \neq \{\mathbf{0}\}$, a **basis** of S is a linearly independent collection of vectors whose span is equal to S. If every vector in S has unit norm, then we call S an **orthonormal basis**. Recall that every basis of a given subspace S has the same number of vectors. This number is called the **dimension** of the subspace S and is denoted by $\dim(S)$. By definition, the dimension of the subspace $\{\mathbf{0}\}$ is zero. The **orthogonal complement** $S^\perp$ of S is defined as

$$S^\perp = \left\{ y \in \mathbb{R}^n : x^T y = 0 \text{ for all } x \in S \right\}.$$

It can be verified that $S^\perp$ is a subspace of $\mathbb{R}^n$, and that if $\dim(S) = k \in \{0, 1, \ldots, n\}$, then we have $\dim(S^\perp) = n - k$. Moreover, we have $S^{\perp\perp} = (S^\perp)^\perp = S$. Finally, every vector $x \in \mathbb{R}^n$ can be uniquely decomposed as $x = x^1 + x^2$, where $x^1 \in S$ and $x^2 \in S^\perp$.

Now, let A be an $m \times n$ real matrix. The **column space** of A is the subspace of $\mathbb{R}^m$ spanned by the columns of A. It is also known as the **range** of A (viewed as a linear transformation $A : \mathbb{R}^n \to \mathbb{R}^m$) and is denoted by

$$\text{range}(A) \equiv \{Ax : x \in \mathbb{R}^n\} \subset \mathbb{R}^m.$$

Similarly, the **row space** of A is the subspace of $\mathbb{R}^n$ spanned by the rows of A. It is well known that the dimension of the column space is equal to the dimension of the row space, and this number is known as the **rank** of the matrix A (denoted by $\text{rank}(A)$). In particular, we have

$$\text{rank}(A) = \dim(\text{range}(A)) = \dim(\text{range}(A^T)).$$

Moreover, we have $\text{rank}(A) \leq \min\{m, n\}$, and if equality holds, then we say that A has **full rank**. The **nullspace** of A is the set $\text{null}(A) \equiv \{x \in \mathbb{R}^n : Ax = \mathbf{0}\}$. It is a subspace of $\mathbb{R}^n$ and has dimension $n - \text{rank}(A)$. The following summarizes the relationships among the subspaces $\text{range}(A)$, $\text{range}(A^T)$, $\text{null}(A)$, and $\text{null}(A^T)$:

$$\begin{aligned} (\text{range}(A))^{\perp} &= \text{null}(A^T), \\ \left(\text{range}(A^T)\right)^{\perp} &= \text{null}(A). \end{aligned}$$

The above implies that given an $m \times n$ real matrix A of rank $r \leq \min\{m, n\}$, we have $\text{rank}(AA^T) = \text{rank}(A^T A) = r$. This fact will be frequently used in the course.

1.5 Affine Subspaces

Let S_0 be a subspace of $\mathbb{R}^n$ and $x^0 \in \mathbb{R}^n$ be an arbitrary vector. Then, the set $S = \{x^0\} + S_0 = \{x + x^0 : x \in S_0\}$ is called an **affine subspace** of $\mathbb{R}^n$, and its dimension is equal to the dimension of the underlying subspace S_0.

Now, let $\mathcal{C} = \{x^1, \ldots, x^m\}$ be a finite collection of vectors in $\mathbb{R}^n$, and let $x^0 \in \mathbb{R}^n$ be arbitrary. By definition, the set $S = \{x^0\} + \text{span}(\mathcal{C})$ is an affine subspace of $\mathbb{R}^n$. Moreover, it is easy to verify that every vector $y \in S$ can be written in the form

$$y = \sum_{i=1}^{m} \left[\alpha_i (x^0 + x^i) + \beta_i (x^0 - x^i)\right]$$

for some $\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_m \in \mathbb{R}$ such that $\sum_{i=1}^{m} (\alpha_i + \beta_i) = 1$; i.e., the vector $y \in \mathbb{R}^n$ is an **affine combination** of the vectors $x^0 \pm x^1, \ldots, x^0 \pm x^m \in \mathbb{R}^n$. Conversely, let $\mathcal{C} = \{x^1, \ldots, x^m\}$ be a finite collection of vectors in $\mathbb{R}^n$, and define

$$S = \left\{ \sum_{i=1}^{m} \alpha_i x^i : \alpha_1, \ldots, \alpha_m \in \mathbb{R}, \sum_{i=1}^{m} \alpha_i = 1 \right\}$$

to be the set of affine combinations of the vectors in $\mathcal{C}$. We claim that S is an affine subspace of $\mathbb{R}^n$. Indeed, it can be readily verified that

$$S = \{x^1\} + \text{span}\left(\{x^2 - x^1, \ldots, x^m - x^1\}\right).$$

This establishes the claim.

Given an arbitrary (i.e., not necessarily finite) collection $\mathcal{C}$ of vectors in $\mathbb{R}^n$, the **affine hull** of $\mathcal{C}$, denoted by $\text{aff}(\mathcal{C})$, is the set of all *finite* affine combinations of the vectors in $\mathcal{C}$. Equivalently, we can define $\text{aff}(\mathcal{C})$ as the intersection of all affine subspaces containing $\mathcal{C}$.

1.6 Some Special Classes of Matrices

The following classes of matrices will be frequently encountered in this course.

- **Invertible Matrix.** An $n \times n$ real matrix A is said to be **invertible** if there exists an $n \times n$ real matrix A^{-1} (called the **inverse** of A) such that $A^{-1}A = I$, or equivalently, $AA^{-1} = I$. Note that the inverse of A is unique whenever it exists. Morever, recall that $A \in \mathbb{R}^{n \times n}$ is invertible iff $\text{rank}(A) = n$.

Now, let A be a non–singular $n \times n$ real matrix. Suppose that A is partitioned as

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

where $A_{ii} \in \mathbb{R}^{n_i \times n_i}$ for $i = 1, 2$, with $n_1 + n_2 = n$. Then, *provided that the relevant inverses exist*, the inverse of A has the following form:

$$A^{-1} = \begin{bmatrix} \left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1} & A_{11}^{-1}A_{12}\left(A_{21}A_{11}^{-1}A_{12} - A_{22}\right)^{-1} \\ \left(A_{21}A_{11}^{-1}A_{12} - A_{22}\right)^{-1}A_{21}A_{11}^{-1} & \left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1} \end{bmatrix}.$$

- **Submatrix of a Matrix.** Let A be an $m \times n$ real matrix. For index sets $\alpha \subset \{1, 2, \ldots, m\}$ and $\beta \subset \{1, 2, \ldots, n\}$, we denote the submatrix that lies in the rows of A indexed by α and the columns indexed by β by $A(\alpha, \beta)$. If $m = n$ and $\alpha = \beta$, the matrix $A(\alpha, \alpha)$ is called a **principal submatrix** of A and is denoted by $A(\alpha)$. The determinant of $A(\alpha)$ is called a **principal minor** of A.

  Now, let A be an $n \times n$ matrix, and let $\alpha \subset \{1, 2, \ldots, n\}$ be an index set such that $A(\alpha)$ is non–singular. We set $\alpha' = \{1, 2, \ldots, n\} \backslash \alpha$. The following is known as the **Schur determinantal formula**:

$$\det(A) = \det(A(\alpha))\det\left[A(\alpha') - A(\alpha', \alpha)A(\alpha)^{-1}A(\alpha, \alpha')\right].$$

- **Orthogonal Matrix.** An $n \times n$ real matrix A is called an **orthogonal matrix** if $AA^T = A^TA = I$. Note that if $A \in \mathbb{R}^{n \times n}$ is an orthogonal matrix, then for any $u, v \in \mathbb{R}^n$, we have $u^Tv = (Au)^T(Av)$; i.e., orthogonal transformations preserve inner products.

- **Positive Semidefinite/Definite Matrix.** An $n \times n$ real matrix A is **positive semidefinite** (resp. **positive definite**) if A is symmetric and for any $x \in \mathbb{R}^n \backslash \{\mathbf{0}\}$, we have $x^TAx \geq 0$ (resp. $x^TAx > 0$). We use $A \succeq \mathbf{0}$ (resp. $A \succ \mathbf{0}$) to denote the fact that A is positive semidefinite (resp. positive definite). We remark that although one can define a notion of positive semidefiniteness for real matrices that are not necessarily symmetric, we shall not pursue that option in this course.

- **Projection Matrix.** An $n \times n$ real matrix A is called a **projection matrix** if $A^2 = A$. Given a projection matrix $A \in \mathbb{R}^{n \times n}$ and a vector $x \in \mathbb{R}^n$, the vector $Ax \in \mathbb{R}^n$ is called the **projection of** $x \in \mathbb{R}^n$ **onto the subspace range**(A). Note that a projection matrix need not be symmetric. As an example, consider

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

  We say that A defines an **orthogonal projection** onto the subspace $S \subset \mathbb{R}^n$ if for every $x = x^1 + x^2 \in \mathbb{R}^n$, where $x^1 \in S$ and $x^2 \in S^\perp$, we have $Ax = x^1$. Note that if A defines an orthogonal projection onto S, then $I - A$ defines an orthogonal projection onto $S^\perp$. Furthermore, it can be shown that A is an orthogonal projection onto S iff A is a symmetric projection matrix with range$(A) = S$.

  As an illustration, consider an $m \times n$ real matrix A, with $m \leq n$ and rank$(A) = m$. Then, the projection matrix corresponding to the orthogonal projection onto the nullspace of A is given by $P_{null(A)} = I - A^T(AA^T)^{-1}A$.

2 Eigenvalues and Eigenvectors

Let A be an $n \times n$ real matrix. We say that $\lambda \in \mathbb{C}$ is an **eigenvalue** of A with corresponding **eigenvector** $u \in \mathbb{C}^n \backslash \{\mathbf{0}\}$ if $Au = \lambda u$. Note that the zero vector $\mathbf{0} \in \mathbb{R}^n$ *cannot* be an eigenvector, although zero *can* be an eigenvalue. Also, recall that given an $n \times n$ real matrix A, there are exactly n eigenvalues (counting multiplicities).

The set of eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$ of an $n \times n$ matrix A is closely related to the **trace** and **determinant** of A (denoted by $\mathrm{tr}(A)$ and $\det(A)$, respectively). Specifically, we have

$$\mathrm{tr}(A) = \sum_{i=1}^{n} \lambda_i \quad \text{and} \quad \det(A) = \prod_{i=1}^{n} \lambda_i.$$

Moreover, we have the following results:

(a) The eigenvalues of A^T are the same as those of A.

(b) For any $c \in \mathbb{R}$, the eigenvalues of $cI + A$ are $c + \lambda_1, \ldots, c + \lambda_n$.

(c) For any integer $k \geq 1$, the eigenvalues of A^k are $\lambda_1^k, \ldots, \lambda_n^k$.

(d) If A is invertible, then the eigenvalues of A^{-1} are $\lambda_1^{-1}, \ldots, \lambda_n^{-1}$.

2.1 Spectral Properties of Real Symmetric Matrices

The **Spectral Theorem for Real Symmetric Matrices** states that an $n \times n$ real matrix A is symmetric iff there exists an orthogonal matrix $U \in \mathbb{R}^{n \times n}$ and a diagonal matrix $\Lambda \in \mathbb{R}^{n \times n}$ such that

$$A = U \Lambda U^T. \tag{2}$$

If the eigenvalues of A are $\lambda_1, \ldots, \lambda_n$, then we can take $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and u^i, the i–th column of U, to be the eigenvector associated with the eigenvalue λ_i for $i = 1, \ldots, n$. In particular, the eigenvalues of a real symmetric matrix are all real, and their associated eigenvectors are orthogonal to each other. Note that (2) can be equivalently written as

$$A = \sum_{i=1}^{n} \lambda_i u^i (u^i)^T,$$

and the rank of A is equal to the number of non–zero eigenvalues.

Note that the set of eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$ of A is unique. Specifically, if $\{\gamma_1, \ldots, \gamma_n\}$ is another set of eigenvalues of A, then there exists a permutation $\pi = (\pi_1, \ldots, \pi_n)$ of $\{1, \ldots, n\}$ such that $\lambda_i = \gamma_{\pi_i}$ for $i = 1, \ldots, n$. This follows from the fact that the eigenvalues of A are the solutions to the polynomial equation

$$\det(A - \lambda I) = 0.$$

On the other hand, the set of unit–norm eigenvectors $\{u^1, \ldots, u^n\}$ of A is not unique. A simple reason is that if u is a unit–norm eigenvector, then $-u$ is also a unit–norm eigenvector. However, there is a deeper reason. Suppose that A has repeated eigenvalues, say, $\lambda_1 = \cdots = \lambda_k = \bar{\lambda}$ for some $k > 1$, with corresponding eigenvectors $u^1, \ldots, u^k$. Then, it can be verified that any vector in the k–dimensional subspace $\bar{\mathcal{L}} = \mathrm{span}\{u^1, \ldots, u^k\}$ is an eigenvector of A with eigenvalue $\bar{\lambda}$. Moreover, all the remaining eigenvectors are orthogonal to $\bar{\mathcal{L}}$. Consequently, each orthonormal basis of $\bar{\mathcal{L}}$

gives rise to a set of k eigenvectors of A whose associated eigenvalue is $\bar{\lambda}$. It is worth noting that if $\{v^1, \ldots, v^k\}$ is an orthonormal basis of $\bar{\mathcal{L}}$, then we can find an orthogonal matrix $P_1^k \in \mathbb{R}^{k \times k}$ such that $V_1^k = U_1^k P_1^k$, where U_1^k (resp. V_1^k) is the $n \times k$ matrix whose i–th column is u^i (resp. v^i), for $i = 1, \ldots, k$. In particular, if $A = U\Lambda U^T = V\Lambda V^T$ are two spectral decompositions of A with

$$\Lambda = \begin{bmatrix} \lambda_{i_1} I_{n_1} & & & \\ & \lambda_{i_2} I_{n_2} & & \\ & & \ddots & \\ & & & \lambda_{i_l} I_{n_l} \end{bmatrix},$$

where $\lambda_{i_1}, \ldots, \lambda_{i_l}$ are the distinct eigenvalues of A, I_k denotes a $k \times k$ identity matrix, and $n_1 + n_2 + \cdots + n_l = n$, then there exists an orthogonal matrix P with the block diagonal structure

$$P = \begin{bmatrix} P_{n_1} & & & \\ & P_{n_2} & & \\ & & \ddots & \\ & & & P_{n_l} \end{bmatrix},$$

where P_{n_j} is an $n_j \times n_j$ orthogonal matrix for $j = 1, \ldots, l$, such that $V = UP$.

Now, suppose that we order the eigenvalues of A as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Then, the **Courant–Fischer theorem** states that the k–th largest eigenvalue λ_k, where $k = 1, \ldots, n$, can be found by solving the following optimization problems:

$$\lambda_k = \min_{w^1, \ldots, w^{k-1} \in \mathbb{R}^n} \max_{\substack{x \neq \mathbf{0}, x \in \mathbb{R}^n \\ x \perp w^1, \ldots, w^{k-1}}} \frac{x^T A x}{x^T x} = \max_{w^1, \ldots, w^{n-k} \in \mathbb{R}^n} \min_{\substack{x \neq \mathbf{0}, x \in \mathbb{R}^n \\ x \perp w^1, \ldots, w^{n-k}}} \frac{x^T A x}{x^T x}. \tag{3}$$

2.2 Properties of Positive Semidefinite Matrices

By definition, a real positive semidefinite matrix is symmetric, and hence it has the properties listed above. However, much more can be said about such matrices. For instance, the following statements are equivalent for an $n \times n$ real symmetric matrix A:

(a) A is positive semidefinite.

(b) All the eigenvalues of A are non–negative.

(c) There exists a unique $n \times n$ positive semidefinite matrix $A^{1/2}$ such that $A = A^{1/2} A^{1/2}$.

(d) There exists an $k \times n$ matrix B, where $k = \text{rank}(A)$, such that $A = B^T B$.

Similarly, the following statements are equivalent for an $n \times n$ real symmetric matrix A:

(a) A is positive definite.

(b) A^{-1} exists and is positive definite.

(c) All the eigenvalues of A are positive.

(d) There exists a unique $n \times n$ positive definite matrix $A^{1/2}$ such that $A = A^{1/2} A^{1/2}$.

Sometimes it would be useful to have a criterion for determining the positive semidefiniteness of a matrix from a block partitioning of the matrix. Here is one such criterion. Let

$$A = \begin{bmatrix} X & Y \\ Y^T & Z \end{bmatrix}$$

be an $n \times n$ real symmetric matrix, where both X and Z are square. Suppose that Z is invertible. Then, the **Schur complement** of the matrix A is defined as the matrix $S_A = X - YZ^{-1}Y^T$. If $Z \succ \mathbf{0}$, then it can be shown that $A \succeq \mathbf{0}$ iff $X \succeq \mathbf{0}$ and $S_A \succeq \mathbf{0}$. There is of course nothing special about the block Z. If X is invertible, then we can similarly define the Schur complement of A as $S'_A = Z - Y^T X^{-1} Y$. If $X \succ \mathbf{0}$, then we have $A \succeq \mathbf{0}$ iff $Z \succeq \mathbf{0}$ and $S'_A \succeq \mathbf{0}$.

3 Singular Values and Singular Vectors

Let A be an $m \times n$ real matrix of rank $r \geq 1$. Then, there exist orthogonal matrices $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ such that

$$A = U \Lambda V^T, \tag{4}$$

where $\Lambda \in \mathbb{R}^{m \times n}$ has $\Lambda_{ij} = 0$ for $i \neq j$ and $\Lambda_{11} \geq \Lambda_{22} \geq \cdots \geq \Lambda_{rr} > \Lambda_{r+1,r+1} = \cdots = \Lambda_{qq} = 0$ with $q = \min\{m, n\}$. The representation (4) is called the **Singular Value Decomposition (SVD)** of A; cf. (2). The entries $\Lambda_{11}, \ldots, \Lambda_{qq}$ are called the **singular values** of A, and the columns of U (resp. V) are called the **left** (resp. **right**) **singular vectors** of A. For notational convenience, we write $\sigma_i \equiv \Lambda_{ii}$ for $i = 1, \ldots, q$. Note that (4) can be equivalently written as

$$A = \sum_{i=1}^{r} \sigma_i u^i (v^i)^T,$$

where u^i (resp. v^i) is the i–th column of the matrix U (resp. V), for $i = 1, \ldots, r$. The rank of A is equal to the number of non–zero singular values.

Now, suppose that we order the singular values of A as $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_q$, where $q = \min\{m, n\}$. Then, the **Courant–Fischer theorem** states that the k–th largest singular value σ_k, where $k = 1, \ldots, q$, can be found by solving the following optimization problems:

$$\sigma_k = \min_{w^1, \ldots, w^{k-1} \in \mathbb{R}^n} \max_{\substack{x \neq \mathbf{0}, x \in \mathbb{R}^n \\ x \perp w^1, \ldots, w^{k-1}}} \frac{\|Ax\|_2}{\|x\|_2} = \max_{w^1, \ldots, w^{n-k} \in \mathbb{R}^n} \min_{\substack{x \neq \mathbf{0}, x \in \mathbb{R}^n \\ x \perp w^1, \ldots, w^{n-k}}} \frac{\|Ax\|_2}{\|x\|_2}. \tag{5}$$

The optimization problems (3) and (5) suggest that singular value and eigenvalue are closely related notions. Indeed, if A is an $m \times n$ real matrix, then

$$\lambda_k(A^T A) = \lambda_k(A A^T) = \sigma_k^2(A) \qquad \text{for } k = 1, \ldots, q,$$

where $q = \min\{m, n\}$. Moreover, the columns of U and V are the eigenvectors of AA^T and $A^T A$, respectively. In particular, our discussion in Section 2.1 implies that the set of singular values of A is unique, but the sets of left and right singular vectors are not. Finally, we note that the largest singular value function induces a matrix norm, which is known as the **spectral norm** and is sometimes denoted by

$$\|A\|_2 = \sigma_1(A).$$

Given an SVD of an $m \times n$ matrix A as in (4), we can define another $n \times m$ matrix $A^\dagger$ by

$$A^\dagger = V \Lambda^\dagger U^T,$$

where $\Lambda^\dagger \in \mathbb{R}^{n \times m}$ has $\Lambda^\dagger_{ij} = 0$ for $i \neq j$ and

$$\Lambda^\dagger_{ii} = \begin{cases} 1/\Lambda_{ii} & \text{for } i = 1, \ldots, r, \\ 0 & \text{otherwise.} \end{cases}$$

The matrices A and $A^\dagger$ possess the following nice properties:

(a) $AA^\dagger$ and $A^\dagger A$ are symmetric.

(b) $AA^\dagger A = A$.

(c) $A^\dagger AA^\dagger = A^\dagger$.

(d) $A^\dagger = A^{-1}$ if A is square and non–singular.

The matrix $A^\dagger$ is known as the **Moore–Penrose generalized inverse** of A. It can be shown that $A^\dagger$ is uniquely determined by the conditions (a)–(c) above.

References

[1] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.

[2] G. Strang. *Introduction to Linear Algebra*. Wellesley–Cambridge Press, Wellesley, Massachusetts, third edition, 2003.

[3] G. Strang. *Linear Algebra and Its Applications*. Brooks/Cole, Boston, Massachusetts, fourth edition, 2006.