MATH10202 Linear Algebra A 2019-20

These notes accompany the part of the course lectured by Mark Kambites.

What is Linear Algebra and Why Do I Need It?

Linear means **"to do with lines"**. **Linear algebra** is the algebra of **linear equations**, which are equations whose solution sets (when drawn in space) are **lines**, and higher dimensional analogues of lines called **linear subspaces**. Linear equations can be concisely and elegantly expressed using algebraic objects called **matrices**, and the first part of the course is mostly concerned with these. The subject is important for both pure mathematics and applications:

- Linear algebra expresses some of the fundamental objects of geometry in a formal algebraic way. It allows us to use equational reasoning to understand geometry.
- Many real-world problems (both those of a geometric flavour and others) are modelled with linear algebra, which provides a powerful toolkit to solve these problems.

Practicalities

Lecturer (first half). Professor Mark Kambites (email `Mark.Kambites@manchester.ac.uk`).

Notes and Lectures. Notes for Mark's part of the course will be provided **with gaps for you to complete in lectures**, on Mark's webpage at

`personalpages.manchester.ac.uk/staff/Mark.Kambites/la.php`

The notes form the definitive content of this part of the course, and you should expect to refer to them frequently. If you need the notes in a different format due a disability, please just let me know. The lectures will explain the same material, but sometimes more informally.

Exercises. Exercise sheets will be handed out in lectures, and contain instructions on when to do the exercises and what to hand in. They are an essential part of learning, so if you want to do well in the course you need to schedule time to attempt all of them (not just the ones for handing in!). Solutions will be made available when you have had a chance to attempt the questions yourself.

Office Hour. My office is **Alan Turing 2.137**. My office hour will generally be **15:30-16:30 on Tuesdays** during teaching weeks; it may sometimes be necessary to change this in which case details will be posted on my website, so please check there before making a special journey. If you can't make it to my office hour but need a personal meeting then please email or ask after a lecture for an appointment.

Supervisions and Homework. The exercise sheets tell you which exercises to hand in for which weeks; your supervision group leader will tell you exactly when and where to hand in. **Attendance at supervisions, and handing in homework, is compulsory.** Please make sure you arrive on time **with the exercise sheets**; group leaders may mark you absent if you come late or not properly prepared.

Assessment. The assessment for the course comprises homework and supervision attendance (10%) and a final exam (90%). The supervisions/homework for the first 6 weeks and Section A of the exam will cover my half of the course.

Feedback. Please let me know how you are finding the course, and especially if you think there are any problems with it. Feel free to speak to me after a lecture, email me, come along in my office hour, or even slip an anonymous note under my office door!

Books. The course is self-contained and full notes will be supplied, so you should not **need** to refer to any books. But if you would like an alternative viewpoint, the course webpage contains some suggested texts.

Prerequisite Material

This course builds directly upon MATH10101 Foundations of Pure Mathematics from last semester. You will need to understand many of the concepts from that course, including:

- proofs (including by contradiction and induction);
- sets (ways of defining them, cardinalities);
- functions (and their properties such as injectivity and surjectivity);
- modular arithmetic.

If you can't remember what the above terms mean, you should go back to your 10101 notes and reread the definitions. Exercise Sheet 0 gives a chance to practice working with these.

Remark. You'll discover this semester that mathematics at university level is **not** an "examine-and-forget" subject. Each stage builds upon what you have learnt before, and demands progressively deeper understanding of previous material. Your main aim in studying this course — and my aim in lecturing it — is therefore **not** the exam, but a deep understanding of the material which will serve you well next year and beyond. If together we can achieve that, the exam will take care of itself!

1 Matrices

Matrices are a useful and elegant way to study linear equations. For teaching purposes it is actually easier to introduce matrices first (which we do in this chapter) and linear equations afterwards (in the next chapter).

Definition. A **matrix** is a rectangular array of real numbers[1], for example:

$$A = \begin{pmatrix} 2 & 5 \\ \sqrt{3} & 0.8 \end{pmatrix} \text{ or } B = \begin{pmatrix} 1 & 6 & 3 \\ 2 & 1 & 9 \end{pmatrix} \text{ or } C = \begin{pmatrix} 1 \\ 2 \\ -3.7 \\ 4 \end{pmatrix}.$$

Size. A matrix has clearly defined numbers of **rows** and **columns**, which together make up its **size**. A matrix with p rows and q columns is called a $p \times q$ **matrix**. Of the above examples:

- A is a 2×2 matrix (2 rows and 2 columns – we say that it is **square**);
- B is a 2×3 matrix (2 rows and 3 columns);
- C is a 4×1 matrix (4 rows and 1 column).

Matrix Entries. If a matrix is called A, we write A_{ij} to denote the entry in the ith row and jth column of A. For example, in the matrix A above, the entries are $A_{11} = 2$, $A_{12} = 5$, $A_{21} = \sqrt{3}$ and $A_{22} = 0.8$. (Some people write a_{ij} instead of A_{ij}.)

[1] For now! Later we shall see that matrices ("**matrices**" is the plural of "matrix") can also have other things in them, such as complex numbers or even more abstract mathematical objects, and much of what we are doing will still work. The objects which form the entries of a matrix are called the **scalars** (so for now "scalar" is just another word for "real number").

Equality of Matrices. Two matrices A and B are **equal** if and only if

- they have the same number of rows **and**
- they have the same number of columns **and**
- corresponding entries are equal, that is $A_{ij} = B_{ij}$ for all appropriate[2] i and j.

1.1 Addition of Matrices

Two matrices can be added if they have the same number of rows (as each other) and the same number of columns (as each other). If so, to get each entry of $A + B$ we just add the corresponding entries of A and B. Formally, $A + B$ is the matrix with entries given by

$$(A+B)_{ij} \;=\; A_{ij} + B_{ij}$$

for all i and j.

Example. The matrices $A = \begin{pmatrix} 2 & 1 \\ 6 & 1 \\ 7 & 1 \end{pmatrix}$ and $B = \begin{pmatrix} 7 & 2 \\ 1 & 4 \\ 3 & 9 \end{pmatrix}$ are both 3×2 matrices, so they can be added. The sum is

$$A + B = \begin{pmatrix} 2+7 & 1+2 \\ 6+1 & 1+4 \\ 7+3 & 1+9 \end{pmatrix} = \begin{pmatrix} 9 & 3 \\ 7 & 5 \\ 10 & 10 \end{pmatrix}.$$

1.2 Scaling of Matrices

Any matrix can be multiplied by any scalar (real number), to give another matrix of the same size. To do this, just multiply each entry of the matrix by the scalar. Formally, if A is a matrix and $\lambda \in \mathbb{R}$ then the matrix λA has entries given by

$$(\lambda A)_{ij} \;=\; \lambda(A_{ij})$$

for all i and j.

Example. $7\begin{pmatrix} -1 & 4 \\ 3 & 2 \end{pmatrix} = \begin{pmatrix} 7 \times (-1) & 7 \times 4 \\ 7 \times 3 & 7 \times 2 \end{pmatrix} = \begin{pmatrix} -7 & 28 \\ 21 & 14 \end{pmatrix}.$

Notation. We write $-A$ as a shorthand for $(-1)A$.

1.3 Subtraction of Matrices

Two matrices of the same size (the same number of rows and same number of columns) can be subtracted. Again, the operation is performed by subtracting corresponding components. Formally,

$$(A-B)_{ij} \;=\; A_{ij} - B_{ij}.$$

Example. Let $A = \begin{pmatrix} 2 & 7 \\ 7 & -1 \\ -1 & 2 \end{pmatrix}$ and $B = \begin{pmatrix} -1 & 3 \\ 0 & 4 \\ 2 & 0 \end{pmatrix}$.

Since both are 3×2 matrices, the subtraction $A - B$ is meaningful and we have

$$A - B = \begin{pmatrix} 2-(-1) & 7-3 \\ 7-0 & -1-4 \\ -1-2 & 2-0 \end{pmatrix} = \begin{pmatrix} 3 & 4 \\ 7 & -5 \\ -3 & 2 \end{pmatrix}$$

[2] By "appropriate" I mean, of course, that i and j have the right values to be "indices" ("**indices**" is the plural of "index") into the matrix: in other words, i is an integer between 1 and the number of rows, and j is an integer between 1 and the number of columns. Once the size of a matrix is known it should be obvious what numbers are allowed to be row and column indices. For brevity I will often just write "for all i and j" to mean "for all integers i between 1 and the number of rows in the matrix and j between 1 and the number of columns in the matrix".

Proposition 1.1. *For any matrices A and B of the same size $p \times q$, we have*

$$A - B \;=\; A + (-B).$$

Proof. Notice first that $(-B) = (-1)B$ is also a $p \times q$ matrix, so the sum $A + (-B)$ makes sense. Now for each i and j we have

$$(A + (-B))_{ij} = A_{ij} + (-B)_{ij} = A_{ij} + (-1)B_{ij} = A_{ij} - B_{ij} = (A - B)_{ij}.$$

So $A - B$ and $A + (-B)$ are the same size, and agree in every entry, which means they are equal.

1.4 Multiplication of Matrices

Two matrices A and B can, **under certain circumstances**, be multiplied together to give a **product** AB. The condition for this to be possible is that the number of **columns in** A should equal the number of **rows in** B.

If A is $m \times n$ and B is $n \times p$ then AB can be formed and will be $m \times p$.

$$\begin{array}{ccccc} A & \times & B & = & AB \\ m \times n & & n \times p & & m \times p \end{array}$$

The entry in the ith row and the jth column of AB depends on the entries in the ith row of A and the entries in the jth column of B. Formally,

$$(AB)_{ij} \;=\; A_{i1}B_{1j} + A_{i2}B_{2j} + A_{i3}B_{3j} + \ldots + A_{in}B_{nj} \;=\; \sum_{k=1}^{n} A_{ik}B_{kj}.$$

The way entries of the factors combine to form entries of the product can be visualised as follows:

$$\underset{\text{i} - \text{th row}}{\begin{pmatrix} \vdots \\ A_{i1}\ A_{i2}\ A_{i3} \ldots A_{in} \\ \vdots \\ \vdots \end{pmatrix}} \underset{\text{j} - \text{th column}}{\begin{pmatrix} & & B_{1j} & \\ & & B_{2j} & \\ \cdots & \cdots & B_{3j} & \cdots \\ & & \vdots & \\ & & B_{nj} & \end{pmatrix}} = \underset{\substack{\text{Element in i} - \text{th row} \\ \text{and j} - \text{th column}}}{\begin{pmatrix} & & \vdots & \\ & & (AB)_{ij} & \\ \cdots & \cdots & \vdots & \cdots \\ & & \vdots & \end{pmatrix}}$$

Example. Consider the matrices

$$A = \begin{pmatrix} 2 & 1 \\ -3 & 7 \\ 1 & 5 \end{pmatrix} \text{ and } B = \begin{pmatrix} 5 \\ -2 \end{pmatrix}.$$

A is 3×2 and B is 2×1. As the number of columns in A is the same as the number of rows in B, AB exists. AB will be 3×1. The three entries of AB are computed as follows:

- The top entry of AB will be computed from top row of A and the single column of B: it is $2 \times 5 + 1 \times (-2) = 8$.
- The middle entry of AB will be computed from the middle row of A and the single column of B: it is $-3 \times 5 + 7 \times (-2) = -29$.
- The bottom entry of AB will be computed from the bottom row of A and the single column of B: it is $1 \times 5 + 5 \times (-2) = -5$.

So $AB = \begin{pmatrix} 8 \\ -29 \\ -5 \end{pmatrix}$.

Example. Consider the matrices

$$C = \begin{pmatrix} 2 & 1 \\ -5 & 3 \end{pmatrix} \text{ and } D = \begin{pmatrix} 4 & -2 \\ 3 & 7 \end{pmatrix}.$$

Both C and D are 2×2. As the number of columns in C equals the number of rows in D, CD exists, and it will also be 2×2.

$$\begin{aligned} CD &= \begin{pmatrix} 2 & 1 \\ -5 & 3 \end{pmatrix} \begin{pmatrix} 4 & -2 \\ 3 & 7 \end{pmatrix} \\ &= \begin{pmatrix} 2 \times 4 + 1 \times 3 & 2 \times (-2) + 1 \times 7 \\ -5 \times 4 + 3 \times 3 & (-5) \times (-2) + 3 \times 7 \end{pmatrix} \\ &= \begin{pmatrix} 11 & 3 \\ -11 & 31 \end{pmatrix} \end{aligned}$$

Example. Consider the matrices

$$E = \begin{pmatrix} 2 & 6 & -4 & 3 \\ 1 & 2 & 8 & -5 \\ -3 & 7 & 2 & 9 \\ 1 & 2 & 2 & 4 \\ -2 & 0 & 7 & 4 \end{pmatrix} \text{ and } F = \begin{pmatrix} 7 & 8 & -1 \\ -1 & 4 & 1 \\ 2 & -3 & 2 \\ 2 & 2 & -2 \end{pmatrix}.$$

Then EF will be a 5×3 matrix, and for example the entry $(EF)_{42}$ can be calculated using the 4th row of E, which is $\begin{pmatrix} 1 & 2 & 2 & 4 \end{pmatrix}$, and the 2nd column of F, which is $\begin{pmatrix} 8 \\ 4 \\ -3 \\ 2 \end{pmatrix}$, so:

$$(EF)_{42} = 1 \times 8 + 2 \times 4 + 2 \times (-3) + 4 \times 2 = 18.$$

Tip. The only way to master matrix operations — especially multiplication — is to **practice**! When you have finished the examples on Exercise Sheet 1, try making up examples of your own and multiplying them. Get a friend to do the same ones and check you get the same answer.

Exercise. Prove that if A and B are **square** matrices (same number of rows and columns) of the same size as each other, then AB is always defined. What size will AB be?

1.5 Properties of Matrix Operations

The operations on matrices share a lot of properties with the corresponding operations on numbers. The following theorem gives some of the most notable ones:

Theorem 1.2. *Let A, B and C be matrices such that the given operations are defined, and λ and μ be numbers. Then*

(i) $A + B = B + A$;

(ii) $A + (B + C) = (A + B) + C$;

(iii) $A(BC) = (AB)C$;

(iv) $A(B+C) = AB + AC$;

(v) $(B+C)A = BA + CA$;

(vi) $A(B-C) = AB - AC$;

(vii) $(B-C)A = BA - CA$;

(viii) $\lambda(B+C) = \lambda B + \lambda C$;

(ix) $\lambda(B-C) = \lambda B - \lambda C$;

(x) $(\lambda+\mu)C = \lambda C + \mu C$;

(xi) $(\lambda-\mu)C = \lambda C - \mu C$;

(xii) $\lambda(\mu C) = (\lambda\mu)C$;

(xiii) $\lambda(BC) = (\lambda B)C = B(\lambda C)$.

Proof. We prove (i) and (iv) to exemplify the methods. Some of the other proofs are on Exercise Sheet 1.

(i) For $A+B$ (and $B+A$) to be defined, A and B must be the same shape, say $m \times n$. Now $A+B$ and $B+A$ are both $m \times n$, and for each i and j we have

$$(A+B)_{ij} = A_{ij} + B_{ij} = B_{ij} + A_{ij} = (B+A)_{ij}.$$

So $A+B$ and $B+A$ are the same shape with exactly the same entries, which means $A+B = B+A$.

(iv) For the given operations to be defined, A must be an $m \times r$ matrix, and B and C both $r \times n$ matrices. Now $A(B+C)$ and $AB+AC$ are both $m \times n$ matrices, and for each i and j we have

$$\begin{aligned}[A(B+C)]_{ij} &= \sum_{l=1}^{r} A_{il}(B+C)_{lj} = \sum_{l=1}^{r} A_{il}(B_{lj}+C_{lj}) = \sum_{l=1}^{r} A_{il}B_{lj} + \sum_{l=1}^{r} A_{il}c_{lj} \\ &= (AB)_{ij} + (AC)_{ij} = (AB+AC)_{ij}.\end{aligned}$$

□

Remark. Property (i) in Theorem 1.2 is called the **commutativity** law for matrix addition; notice that there isn't a corresponding statement for matrix multiplication - see Exercise Sheet 1. Properties (ii) and (iii) are **associativity** laws (for matrix addition and matrix multiplication respectively). Properties (iii)-(xi) are **distributivity** laws.

1.6 Diagonal Matrices, Identities and Inverses

The Main Diagonal. If A is a square matrix (say $n \times n$) then **the main diagonal** (sometimes just called **the diagonal**) is the diagonal from top left to bottom right of the matrix. In other words, it is the collection of entries whose row number is the same as their column number.

A matrix is called **diagonal** if all the entries **not** on the diagonal are 0. (Some, or even all, of the diagonal entries may also be 0.) For example,

$$\begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

are diagonal matrices but

$$\begin{pmatrix} -2 & 0 \\ 5 & 2 \end{pmatrix}$$

is not diagonal because of the 5 which lies off the main diagonal.

Identity Matrices. A square matrix with 1's on the main diagonal and 0's everywhere else is called an **identity matrix**. The $n \times n$ identity matrix is denoted I_n. For example:

$$I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ while } I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

It has the important property that

$$AI_n = A \text{ and/or } I_n A = A$$

whenever A is the right size for these products to be defined. (**Exercise:** prove this.)

Inverses. Let A be an $n \times n$ square matrix. An **inverse** of A is an $n \times n$ matrix A^{-1} such that

$$A\,A^{-1} \;=\; A^{-1}\,A \;=\; I_n.$$

Not every matrix has an inverse. Matrices with an inverse are called **invertible**; those which are not invertible are called **singular**.

Exercise. What is the inverse of the identity matrix I_n?

Lemma 1.3. *(i) Suppose A is an invertible $n \times n$ matrix. Then the inverse of A is unique.*

(ii) If E and D are both invertible $n \times n$ matrices then ED is invertible.

Proof. (i) Suppose B and C are both inverses of A. Then $AB = BA = I_n$ and $AC = CA = I_n$. Now using Theorem 1.2(iii):

$$B \;=\; I_n B \;=\; (CA)B \;=\; C(AB) \;=\; CI_n \;=\; C.$$

(ii) Let E^{-1} and D^{-1} be inverses of E and D respectively. Then again by Theorem 1.2(iii):

$$(ED)(D^{-1}E^{-1}) = E(DD^{-1})E^{-1} = EI_nE^{-1} = EE^{-1} = I_n.$$

A similar argument (**exercise!**) gives $(D^{-1}E^{-1})(ED) = I_n$. This shows that the matrix $D^{-1}E^{-1}$ is an inverse of ED (and so by part (i), **the** inverse of ED).

□

Notation. If A is an invertible matrix, we write A^{-1} for its inverse.

There is a simple way to describe inverses of 2×2 matrices:

Theorem 1.4. *Let*

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

be a 2×2 matrix. Then A is invertible if and only if $ad - bc \neq 0$, and if it is, then

$$A^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Remark. The value $ad - bc$ is a very important number associated to the matrix A. It is called the **determinant** of A, and we will see later (Chapter 4) how to generalise the idea to bigger matrices.

Exercise. Determine which of the following matrices are invertible, and for those which are find the inverse:

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \begin{pmatrix} -1 & 3 \\ -2 & 6 \end{pmatrix} \quad \begin{pmatrix} 3 & 1 \\ 3 & 1 \end{pmatrix}.$$

The Zero Matrix. A matrix is called a **zero matrix** if all of its entries are 0. For example:

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \text{ or } \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Clearly there is exactly one zero matrix of each size: we write $\mathbf{0}_{p\times q}$ for the $p \times q$ matrix all of whose entries are 0, so those above are $\mathbf{0}_{4\times 1}$ and $\mathbf{0}_{2\times 2}$.

1.7 Transpose and Symmetry

The **transpose** of a $p \times q$ matrix is the $q \times p$ matrix obtained by swapping over the rows and columns. The transpose of a matrix A is written A^T, and formally it is the $q \times p$ matrix which entries given by

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

for all i and j. For example:

$$\begin{pmatrix} 1 \\ 7 \\ 2 \\ 4 \end{pmatrix}^T = \begin{pmatrix} 1 & 7 & 2 & 4 \end{pmatrix} \text{ while if } A = \begin{pmatrix} 2 & 6 \\ -1 & 3 \\ 1 & 1 \end{pmatrix} \text{ then } A^T = \begin{pmatrix} 2 & -1 & 1 \\ 6 & 3 & 1 \end{pmatrix}.$$

Theorem 1.5. *Let A and B be matrices such that the given operations are defined. Then:*

(i) $(A^T)^T = A$;

(ii) $(A + B)^T = A^T + B^T$;

(iii) $(AB)^T = (B^T A^T)$.

Proof.

(i) Clear.

(ii) For $A + B$ to be defined we need A and B to have the same size, say $m \times n$, in which case $(A + B)^T$ and $A^T + B^T$ are both defined and $n \times m$. Now for each i and j,

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

So $(A + B)^T$ and $(A^T + B^T)$ are the same size with the same entries, which means they are equal.

(iii) For AB to be defined it must be that A is $m \times p$ and B is $p \times n$, in which case A^T is $p \times m$ and B^T is $n \times p$, so that $B^T A^T$ is defined. It is easy to check that $(AB)^T$ and $B^T A^T$ are both $n \times m$. Now for $1 \le i \le n$ and $1 \le j \le m$ we have[3]

$$((AB)^T)_{ij} = (AB)_{ji} = \sum_{k=1}^{p} A_{jk} B_{ki} = \sum_{k=1}^{p} (A^T)_{kj} (B^T)_{ik} = \sum_{k=1}^{p} (B^T)_{ik} (A^T)_{kj} = (B^T A^T)_{ij}.$$

□

[3] This is a case where just saying "appropriate i and j" might not be entirely clear, so we are careful to specify!

Theorem 1.6. *If A is an invertible matrix then A^T is invertible and $(A^T)^{-1} = (A^{-1})^T$.*

Proof. Using Theorem 1.5(iii)

$$A^T(A^{-1})^T = (A^{-1}A)^T = (I_n)^T = I_n$$
$$(A^{-1})^T A^T = (AA^{-1})^T = (I_n)^T = I_n$$

□

Definition. A matrix A is called **symmetric** if $A^T = A$, or **skew symmetric** if $A^T = -A$.

Example. $\begin{pmatrix} 1 & 9 & 4 \\ 9 & 3 & -2 \\ 4 & -2 & 7 \end{pmatrix}$ is symmetric. $\begin{pmatrix} 0 & 9 & 4 \\ -9 & 0 & -2 \\ -4 & 2 & 0 \end{pmatrix}$ is skew symmetric.

Remark. Notice that a matrix which is symmetric or skew symmetric must be square, and a skew symmetric matrix must have 0s on the main diagonal.

Theorem 1.7. *(i) If A is any matrix then AA^T and A^TA are symmetric matrices.*

(ii) If A is a square matrix then $A + A^T$ is symmetric and $A - A^T$ is skew symmetric.

(iii) If A is invertible and symmetric then A^{-1} is symmetric

Proof. (i) and (ii) follow easily from Theorem 1.5, and (iii) follows from Theorem 1.6. (**exercise**: do this!). □

2 Systems of Linear Equations

A system of equations of the form

$$\begin{array}{rcrcr} x & + & 2y & = & 7 \\ 2x & - & y & = & 4 \end{array}$$

or

$$\begin{array}{rcrcrcr} 5p & - & 6q & + & r & = & 4 \\ 2p & + & 3q & - & 5r & = & 7 \\ 6p & - & q & + & 4r & = & -2 \end{array}$$

is called a **system of linear equations**.

Definition. An equation involving certain variables is **linear** if each side of the equation is a sum of **constants** and **constant multiples of the variables**.

Examples. The equations $2x = 3y + \sqrt{3}$, $3 = 0$ and $x + y + z = 0$ are linear.

Non-examples. The equations $y = x^2$ and $3xy + z + 7 = 2$ are **not** linear since they involve **products** of variables.

2.1 Solutions to Systems of Equations

A **solution** to a system of equations is a way of assigning a value to each variable, so as to make all the equations true at once.

Example. The system of equations ...

$$\begin{array}{rcrcr} 4x & - & 2y & = & 22 \\ 9x & + & 3y & = & 12 \end{array}$$

... has a solution $x = 3$ and $y = -5$. (In fact in this case this is the **only** solution.)

Remark. In general, a system of linear equations can have **no** solutions, a **unique** solution, or **infinitely many** solutions. It is **not** possible for the system to have, for example, exactly 2 solutions; this is a higher dimensional generalisation of the fact that 2 straight lines cannot meet in exactly 2 points (recall Exercise Sheet 0).

Definition. The set of all solutions is called the **solution set** or **solution space** of the system.

Warning. When we talk about **solving** a system of equations, we mean describing **all** the solutions (or saying that there are none, if the solution set is empty), not just finding one possible solution!

2.2 2-dimensional Geometry of Linear Equations

What can solution spaces look like? Consider a linear equation in two variables x and y, say:

$$3x + 4y = 4.$$

Each solution consists of an x-value and a y-value. We can think of these as the coordinates of a point in 2-dimensional space. For example, a solution to the above equation is $x = -4$ and $y = 4$. This solution gives the point $(-4, 4)$.

In general, the set of solutions to a linear equation with variables x and y forms a **line** in 2-dimensional space. So each equation corresponds to a line — we say that the equation **defines** the line.

Systems of Equations. Now suppose we have a **system** of two such equations. A solution to the system means an x-value and a y-value which solve **all the equations at once**. In other words, a solution corresponds to a point which lies on **all** the lines. So in geometric terms, the solution set to a system of linear equations with variables x and y is an **intersection of lines** in the plane.

Question. Recall from Exercise Sheet 0 what an intersection of two lines in the plane can look like:

- Usually[4] it is a **single point**. This is why 2 equations in 2 variables usually have a unique solution.
- Alternatively, the lines could be distinct but parallel. Then the intersection is the **empty set**. This is why 2 equations in 2 variables can sometimes have no solutions.
- Or then again, the two lines could actually be the same. In this case the intersection is the **whole line**. This is why 2 equations in 2 variables can sometimes have infinitely many solutions.

Similarly, if we have a system of **more than** 2 equations, we are looking for the intersection of all the lines (in other words, the set of points in the plane which lie on all the lines). The intersection of 3 lines is "usually" empty but could also be a single point or a line.

2.3 Higher Dimensional Geometry of Linear Equations

Now suppose we have an equation in the variables x, y and z, say

$$3x + 4y + 7z = 4.$$

This time we can consider a solution as a point in 3-dimensional space. The set of all solutions (assuming there are some) forms a **plane** in space. The solution set for a **system** of k linear equations will be an **intersection of k planes**.

Exercise. What can the intersection of 2 planes in 3-space look like? Try to imagine all the possibilities, as we did for lines in the previous section. How do they correspond to the possible forms of solution sets of 2 equations with 3 variables?

Exercise. Now try to do the same for intersections of 3 planes. (Recall that the intersection of three sets is the set of points which lie in all three). How do the possible geometric things you get correspond to the possible forms of solution sets for 3 equations with 3 variables?

In yet higher dimensions, the solution set to a linear equation in n variables is a copy of $(n-1)$-dimensional space inside n-dimensional space. Such a set is called a **hyperplane**. So geometrically, the solution set of k equations with n variables will be an **intersection of k hyperplanes** in n-dimensional space.

Remark. Actually it is not **quite** true to say that every equation defines a hyperplane. Equations like "$0 = 0$" and "$2 = 2$" define the whole space; if **all** equations in a system have this form then the solution space will be the whole space. On the other hand, equations like "$0 = 1$" have no solutions; if **any** equation in a system has this form then the solution space will be empty.

Remark. You might be wondering what happens if we drop the requirement that the equations be linear, and allow for example **polynomial** equations? If we do this then lines, planes and hyperplanes get replaced by more general objects called **curves**, **surfaces** and **hypersurfaces** respectively. For example, the solution set to $x^2 + y^2 + z^2 = 1$ is a sphere, which is a 2-dimensional surface in 3-space. The solution space to a **system** of polynomial equations (an intersection of hypersurfaces) is called an **(affine) algebraic variety**. This part of mathematics is called **algebraic geometry**; you can study it in your third year.

[4] Of course "usually" here is a vague term, but I hope you can see intuitively what I mean! If you pick two lines at random it is somehow "infinitely unlikely" that they will be exactly parallel. It is possible to make this intuition precise; this is well beyond the scope of this course but as a (challenging!) exercise you might like to think how you would do so.

2.4 Linear Equations and Matrices

Any system of linear equation can be rearranged to put all the constant terms on the right, and all the terms involving variables on the left. This will yield a system something like ...

$$\begin{array}{rcrcl} 4x & - & 2y & = & 3 \\ 2x & + & 2y & = & 5 \end{array}$$

... so from now on we will assume all systems have this form. The system can then be written in **matrix form**:

$$\begin{pmatrix} 4 & -2 \\ 2 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3 \\ 5 \end{pmatrix}.$$

You can check (by multiplying out the matrix equation using the definition of matrix multiplication) that it holds if and only if both of the original (scalar) equations hold. We can express the equations even more concisely as an **augmented matrix**:

$$\left(\begin{array}{cc|c} 4 & -2 & 3 \\ 2 & 2 & 5 \end{array}\right).$$

2.5 Elementary Row Operations and Solving Systems of Equations

At school you learnt to solve systems of linear equations by simple algebraic manipulation. This *ad hoc* approach is handy for small examples, but many applications involve bigger systems of equations, and for these it is helpful to have a **systematic method**.

We start with the augmented matrix of our system (see Section 2.4), and we allow ourselves to **modify** the matrix by certain kinds of steps. We can:

(i) **multiply** a row by a **non-zero** scalar (written $r_i \to \lambda r_i$);

(ii) **add** a multiple of one row to another (written $r_i \to r_i + \lambda r_j$);

(iii) **swap** two rows of the matrix (written $r_i \leftrightarrow r_j$).

These are called **elementary row operations**.[5]

Warning. Multiplying a row by 0 is definitely **not** an elementary row operation!

Definition. Two matrices A and B are called **row equivalent** if each can be obtained from the other by a sequence of elementary row operations.

Row operations and row equivalence are useful because of the following fact:

Theorem 2.1. *Suppose M and N are the augmented matrices of two system of linear equations. If M is row equivalent to N then the two systems have the same solution set.*

We'll prove this theorem later (Chapter 3), and but for now let's explore its consequences. It means that if we start with a system of equations, applying elementary row operations to the augmented matrix will give us another system of equations **with the same solution set**. The aim of the game is to obtain a system of equations which is easier to solve. For example, consider the system:

First we convert this into augmented matrix form (noting that the "missing" x terms are really $0x$ and become 0s in the matrix):

[5] There is an obvious "dual" idea of **elementary column operations**, written $c_i \to \lambda c_i$, $c_i \to c_i + \lambda c_j$ and $c_i \leftrightarrow c_j$.

Now we swap rows 1 and 2 ($r_1 \leftrightarrow r_2$), to get

Next let's scale rows 1 and 2 to make the first entry in each 1 ($r_1 \to -r_1,\ r_2 \to \frac{1}{10}r_2$):

Next we subtract 10 times row 2 from row 3 ($r_3 \to r_3 - 10r_2$), giving:

Now we convert the augmented matrix back to a system of equations:

It turns out that these equations are easier to solve than the original ones. Equation [3] tells us straight away that $z = 3$. Now substituting this into equation [2] we get $y - \frac{3}{10} \times 3 = -\frac{19}{10}$, in other words, $y = -1$. Finally, substituting both of these into equation [2] gives $x - 4 \times (-1) + 3 = 9$, that is, $x = 2$.

So we have found the solutions to equations [1], [2] and [3]. But these were obtained from our original system of equations by elementary row operations, so (by Theorem 2.1) they are also the solutions to the original system of equations. (**Check this for yourself!**)

This strategy forms the basis of a very general algorithm, called **Gaussian**[6] **Elimination**, which we shall see in Section 2.7.

2.6 Row Echelon Matrices

Remark. Notice how we converted the matrix into a roughly "triangular" form, with all the entries towards the "bottom left" being 0. It was this property which made the new equations easy to solve. The following definitions will make precise what we mean by "triangular" here.

Pivots. The **pivot** (or **leading entry**) of a row in an augmented matrix is the position of the leftmost non-0 entry to the left of the bar. (If all entries left of the bar are 0, we call the row a **zero row** and it has no pivot.)

Example. In the matrix A on the right

$$A = \left(\begin{array}{ccc|c} 0 & 5 & 1 & 0 \\ 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 \end{array}\right)$$

- the pivot of the first row is the 5;
- the pivot of the second row is the 3;
- the third row is a zero row (it has no pivot)

Row Echelon Matrices. An augmented matrix is called a **row echelon matrix** if

[6]After Carl Friedrich Gauß (pronounced to rhyme with "house"), 1777–1855.

(1) any zero rows come at the bottom; and

(2) the pivot of each of the other rows is **strictly to the right of the pivot of the row above**; and

(3) all the pivots are 1.

A row echelon matrix is a **reduced row echelon matrix** if in addition:

(4) each pivot is the only non-zero entry in its column.

Example. The matrix A above is **not** a row echelon matrix, because the second row pivot is not to the right of the first row pivot. The matrix B on the right **is** a row echelon matrix, but it is **not** a **reduced** row echelon matrix because the 2 in the first row is in the same column as the second row pivot.

$$B = \left(\begin{array}{cccc|c} 0 & 1 & 1 & 2 & 0 \\ 0 & 0 & 0 & 1 & 5 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 6 \end{array}\right)$$

2.7 Gaussian Elimination

Gaussian Elimination is a simple algorithm which **uses elementary row operations to reduce the augmented matrix to row echelon form**:

(1) By swapping rows if necessary, make sure that **no non-0 entries (in any row) are to the left of the first row pivot.**

(2) Scale the first row so that the pivot is 1.

(3) Add multiples of the first row to each of the rows below, so as to make all entries below the first row pivot become 0. (Since there were no non-0 entries left of the first row pivot, this ensures all pivots **below** the first row are strictly to the right of the pivot **in** the first row.)

(4) Now ignore the first row, and repeat the entire process with the second row. (This will ensure that all pivots **below** the second row are to the right of the pivot **in** the second row, which in turn is to the right of the pivot in the first row.)

(5) Keep going like this (with the third row, and so on) until either we have done all the rows, or all the remaining rows are zero.

(6) At this point the matrix is in row echelon form.

Exercise. Go back to the example in Section 2.5, and compare what we did with the procedure above.

2.8 Another Example

Example. Let's solve the system of equations:

$$\begin{array}{rcrcrcr} 2x & + & y & - & z & = & -7 \\ 6x & & & - & z & = & -10 \\ -4x & + & y & + & 7z & = & 31 \end{array}$$

Solution. First we express the system as an augmented matrix:

Notice (Step 1) that there are no non-0 entries to the left of the pivot in the first row (there can't be, since the pivot is in the first column!). So we do not need to swap rows to ensure this is the case.

Next (Step 2) we scale to make the pivot in the first row 1, that is, $r_1 \to \frac{1}{2}r_1$, which gives

Next (Step 3) we seek to remove the entry in row 2 which is below the row 1 pivot. We can do this with the operation $r_2 \to r_2 - 6r_1$, giving the matrix:

Similarly, we remove the entry in row 3 below the row 1 pivot by $r_3 \to r_3 + 4r_1$:

Now we ignore the first row, and repeat the whole process with the remaining rows. Notice that there are no pivots to the left of the row 2 leading entry. (The one in row 1 doesn't count, since we are ignoring row 1!). Now we scale to make the second row pivot into 1, with $r_2 \to -\frac{1}{3}r_2$, giving:

Then remove the entry in row 3 which is below the row 2 pivot, with $r_3 \to r_3 - 3r_2$, giving:

Finally, we repeat the process again, this time ignoring rows 1 and 2. All we have to do now is scale row 3 to make the pivot 1, with $r_3 \to \frac{1}{7}r_3$:

Our matrix is now in row echelon form, so we convert it back to a system of equations:

These matrices can easily be solved by "backtracking" through them. Specifically:

- equation [3] says explicitly that $z = 4$;
- substituting into equation [2] gives $y - \frac{2}{3} \times 4 = -\frac{11}{3}$, so $y = -1$;
- substituting into equation [1] gives $x + \frac{1}{2} \times (-1) - \frac{1}{2} \times 4 = -\frac{7}{2}$, so $x = -1$.

Thus, the solution is $x = -1$, $y = -1$ and $z = 4$. (You should **check this** by substituting the values back into the original equations!)

Remark. Notice how the row echelon form of the matrix facilitated the "backtracking" procedure for solving equations [1], [2] and [3].

Exercise. Use Gaussian elimination to solve the following systems of equations:

$$(i)\ \begin{array}{rcrcr} 4x & + & y & = & 9 \\ 2x & - & 3y & = & 1 \end{array} \qquad (ii)\ \begin{array}{rcrcr} 2x & - & 4y & = & 12 \\ -x & + & 2y & = & -5 \end{array}$$

2.9 Gauss-Jordan Elimination

Gaussian elimination lets us transform any matrix into a row echelon matrix. Sometimes it is helpful to go further and obtain a **reduced** row echelon matrix. To do this we use a slight variation called **Gauss-Jordan**[7] **Elimination**:

- First use Gaussian elimination to compute a row echelon form.
- Add multiples of the 2nd row to the row above to remove any non-0 entries above the 2nd row pivot.
- Add multiples of the 3rd row to the rows above to remove any non-0 entries above the 3rd row pivot.
- Continue down the rows, adding multiples of row k to the rows above to eliminate any non-0 entries above the kth row pivot.

Theorem 2.2. *Let A be a matrix. Then A is row equivalent to a* ***unique*** *reduced row echelon matrix (called the* ***reduced row echelon form of the matrix****).*

Proof. It should be clear that applying the Gauss-Jordan elimination algorithm to A will give a reduced row echelon matrix row equivalent to A. The proof of uniqueness is more difficult, and we omit it. ☐

Important Consequence. We can check if two matrices are row equivalent, by applying Gauss-Jordan elimination to compute their reduced row echelon forms, then checking if these are equal. (If the reduced row echelon forms are equal then clearly the original matrices are equivalent. Conversely, if the matrices are equivalent then by Theorem 2.2, they must have the same reduced row echelon form.)

Exercise. Check if $\begin{pmatrix} 1 & 3 & 6 \\ 0 & 1 & 2 \\ 3 & 7 & 14 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 & 0 \\ 1 & 5 & 10 \\ 1 & 1 & 2 \end{pmatrix}$ are row equivalent.

2.10 Elimination with More Variables than Equations

What if, after applying Gaussian elimination and converting back to equations, there are more variables than equations? This can happen either because the system we started with was like this, or because elimination gives us rows of zeros in the matrices: these correspond to the equation "$0 = 0$" which obviously can't be used to find the value of any variables.

Suppose there are k equations and n variables, where $n > k$. We can still apply Gaussian elimination to find a row echelon form. But when we convert back to equations and "backtrack" to find solutions, we will sometimes still encounter an equation with more than one unknown variable.

If this happens, there are **infinitely many** solutions. We can describe all the solutions by introducing **parameters** to replace $n - k$ of the variables.

[7] After the German Wilhelm Jordan (1842–1899) and **not** the more famous French mathematician Camille Jordan (1838–1922). The correct pronunciation is therefore something like "YOR-dan" (not zhor-DAN!) but most people give up and go with the English pronunciation!

Example. Consider the equations.

$$\begin{array}{rcrcrcr} -x & + & 2y & + & z & = & 2 \\ 3x & - & 5y & - & 2z & = & 10 \\ x & - & y & & & = & 14 \end{array}$$

The augmented matrix is:

$$\left(\begin{array}{ccc|c} -1 & 2 & 1 & 2 \\ 3 & -5 & -2 & 10 \\ 1 & -1 & 0 & 14 \end{array}\right) \text{ and reducing to row echelon form we get } \left(\begin{array}{ccc|c} 1 & -2 & -1 & -2 \\ 0 & 1 & 1 & 16 \\ 0 & 0 & 0 & 0 \end{array}\right).$$

which gives the equations

$$\begin{array}{rcrcrcrl} x & - & 2y & - & z & = & -2 & [1] \\ & & y & + & z & = & 16 & [2] \\ & & & & 0 & = & 0 & [3] \end{array}$$

Equation [3] is always satisfied but clearly useless for finding the values of the variables, so we can ignore it. Equation [2] has two unknown variables (y and z) so we introduce a parameter λ to stand for one of them, say $z = \lambda$. Now solving we get $y = 16 - \lambda$. Substituting into [1] gives $x - 2(16 - \lambda) - \lambda = -2$, or $x = 30 - \lambda$. So the solutions are:

$$x = 30 - \lambda, \quad y = 16 - \lambda, \quad z = \lambda$$

for $\lambda \in \mathbb{R}$.

Remark. When we say these are the solutions, we mean that substituting in different values of $\lambda \in \mathbb{R}$ will give all the solutions to the equations. For example, $\lambda = 1$ would give $x = 29$, $y = 15$ and $z = 1$, so this is one solution to the system of equations (**check this!**). Or then again, $\lambda = 0$ gives $x = 30$, $y = 16$ and $z = 0$, so this is another possible solution.

Exercise. Use Gaussian elimination to solve

$$\begin{array}{rcrcr} 2x & - & 4y & = & 10 \\ -x & + & 2y & = & -5 \end{array}$$

2.11 Homogeneous Systems of Equations

A system of linear equations is called **homogeneous** if the constant terms are all 0. For example:

$$\begin{array}{rcrcrcr} x & + & 7y & - & 4z & = & 0 \\ 2x & + & 4y & - & z & = & 0 \\ 3x & + & y & + & 2z & = & 0 \end{array}$$

An obvious observation about homogeneous systems is that they **always have at least one solution**, given by setting all the variables equal to 0. This is called the **trivial solution**. Geometrically, this means that the solution space to a homogeneous system always contains the origin.

Forward Pointer. Solution sets to homogeneous systems — that is, intersections of hyperplanes through the origin — are very special and important subsets in $\mathbb{R}^n$, called **vector subspaces**. More on this later (Chapter 5).

When it comes to solving them, homogeneous systems of equations are treated just like non-homogeneous ones. The only difference is that nothing interesting ever happens on the RHS of the augmented matrix, as the following exercise will convince you!

Exercise. Find solutions to the system of equations above.

Remark. Notice how, in your solution, everything stays "homogeneous" throughout. All the augmented matrices in the elimination process, and all the resulting equations, have only 0's on the RHS.

2.12 Computation in the Real World

In describing methods to solve equations (Gaussian and Gauss-Jordan elimination) we have implicitly assumed two things:

- that the equations are precisely known; and
- that we can perform the necessary arithmetic with perfect accuracy.

But in reality, these assumptions are often not true. If we want to solve a system of equations because it models a real-world problem, then the coefficients in the equations may have been obtained by measurement, in which case their values are subject to experimental error. And even if the coefficients are known precisely, real-world computers don't actually store and process real numbers: they can't, because there is an (uncountably) infinite set of real numbers but even the most powerful digital computer has a finite memory. Instead they work with **approximations** to real numbers: typical is **floating point arithmetic** in which numbers are rounded to a certain number of significant figures. At every stage of the computation numbers have to be rounded to the nearest available approximation, introducing slight inaccuracies even if the original data was perfect.

If the starting data or the arithmetic are imperfect then it clearly isn't realistic to expect perfect solutions. But some methods have the useful property that data which is **almost** right will produce solutions which are **almost** right[8]. Other methods do not have this property at all: they may work perfectly in theory, but the tiniest error in the starting data gets magnified as the computation progresses, and leads to a wildly inaccurate answer. Gaussian elimination, in the simple form we have described, is not very good in this respect, but it can be made better with a slightly more sophisticated approach (called **partial pivoting**) which takes account of the relative size of the matrix entries.

Further consideration of these practical issues is beyond the scope of this course, but you can learn about them in later years, starting with the second year option **MATH20602 Numerical Analysis I**.

[8] Or the subtly different property that the solutions produced are exact solutions to **almost the right problems**. Of course this is an oversimplification; for an accurate explanation see the first few sections of Chapter 1 in the book *Accuracy and Stability of Numerical Algorithms (2nd edition)* by N. J. Higham, and in particular the distinction between forward error and backward error. (But this is not examinable in this course.)