Chapter 5. Matrices

This material is in Chapter 1 of Anton & Rorres.

5.1 Basic matrix notation

We recall that a *matrix* is a rectangular array or table of numbers. We call the individual numbers *entries* of the matrix and refer to them by their row and column numbers. The rows are numbered $1, 2, \ldots$ from the top and the columns are numbered $1, 2, \ldots$ from left to right.

In the example

$$\begin{bmatrix} 1 & 1 & 2 & 5 \\ 1 & 11 & 13 & -2 \\ 2 & 1 & 3 & 4 \end{bmatrix}$$

13 is the $(2, 3)$ entry, the entry in row 2 and column 3.

Now for some terminology we did not discuss before.

The matrix above is called a 3×4 matrix because it has 3 rows and 4 columns. We can talk about matrices of all different sizes such as

$$\underset{2\times 2}{\begin{bmatrix} 4 & 5 \\ 7 & 11 \end{bmatrix}} \qquad \underset{2\times 1}{\begin{bmatrix} 4 \\ 7 \end{bmatrix}} \qquad \underset{1\times 2}{\begin{bmatrix} 4 & 7 \end{bmatrix}} \qquad \underset{3\times 2}{\begin{bmatrix} 4 & 5 \\ 7 & 11 \\ 13 & 13 \end{bmatrix}}$$

and in general we can have $m \times n$ matrices for any $m \geq 1$ and $n \geq 1$.

Matrices with just one row are called *row matrices*. A $1 \times n$ matrix $\begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$ has just the same information in it as an n-tuple $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ and so we could be tempted to identify $1 \times n$ matrices with n-tuples (which we know are points or vectors in $\mathbb{R}^n$).

We use the term *column matrix* for a matrix with just one column. Here is an $n \times 1$ (column) matrix

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

and again it is tempting to think of these as the "same" as n-tuples $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$. Maybe not quite as tempting as it is for row matrices, but not such a very different idea.

To avoid confusion that would certainly arise if we were to make either of these identifications (either of $1 \times n$ matrices with n-tuples *or* of $n \times 1$ matrices with n-tuples) we will not make either of them and keep all the different objects in their own separate places. A bit later on, it will often be more convenient to think of coluumn $n \times 1$ matrices as points of $\mathbb{R}^n$, but we will not come to that for some time.

Now, to clarify any confusion these remarks might cause, we explain that we consider two matrices to be the 'same' matrix only if they are absolutely identical. They have to have the

same shape (same number of rows and same number of columns) and they have to have the same numbers in the same positions. Thus, all the following are different matrices

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \neq \begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix} \neq \begin{bmatrix} 2 & 1 & 0 \\ 3 & 4 & 0 \end{bmatrix} \quad \begin{bmatrix} 2 & 1 \\ 3 & 4 \\ 0 & 0 \end{bmatrix}$$

5.2 Double subscripts

When we want to discuss a matrix without listing the numbers in it, that is when we want to discuss a matrix that is not yet specified or an unknown matrix we use a notation like this with double subscripts

$$\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}$$

This is a 2×2 matrix where the $(1,1)$ entry is x_{11}, the $(1,2)$ entry is x_{12} and so on.

It would probably be clearer of we put commas in and write

$$\begin{bmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & x_{2,2} \end{bmatrix}$$

instead, but people commonly use the version without the commas between the two subscripts.

Carrying this idea further, when we want to discuss an $m \times n$ matrix x and refer to its entries we write

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

So the (i,j) entry is called x_{ij}.

Sometimes we want to write something like this but we don't want to take up space for the whole picture and we write an abberviated version like

$$x = [x_{ij}]_{1 \leq i \leq m, 1 \leq j \leq n}$$

To repeat what we said about when matrices are equal using this kind of notation, suppose we have two $m \times n$ matrices

$$x = [x_{ij}]_{1 \leq i \leq m, 1 \leq j \leq n} \text{ and } y = [y_{ij}]_{1 \leq i \leq m, 1 \leq j \leq n}$$

Then $x = y$ means the mn scalar equations $x_{ij} = y_{ij}$ must all hold (for each (i,j) with $1 \leq i \leq m, 1 \leq j \leq n$). And if an $m \times n$ matrix equals an $r \times s$ matrix, we have to have $m = r$ (same number or rows), $n = s$ (same number of columns) and then all the entries equal.

5.3 Arithmetic with matrices

In much the same way as we did with n-tuples we now define addition of matrices. We only allow addition of matrices that are of the same size. Two matrices of different sizes cannot be added.

If we take two $m \times n$ matrices

$$x = [x_{ij}]_{1 \leq i \leq m, 1 \leq j \leq n} \text{ and } y = [y_{ij}]_{1 \leq i \leq m, 1 \leq j \leq n}$$

then we define

$$x + y = [x_{ij} + y_{ij}]_{1 \leq i \leq m, 1 \leq j \leq n}$$

(the $m \times n$ matrix with $(1,1)$ entry the sum of the $(1,1)$ entries of x and y, $(1,2)$ entry the sum of the $(1,2)$ entries of x and y, and so on).

For example

$$\begin{bmatrix} 2 & 1 \\ 3 & -4 \\ 0 & 7 \end{bmatrix} + \begin{bmatrix} 6 & -2 \\ 15 & 12 \\ -9 & 21 \end{bmatrix} = \begin{bmatrix} 2+6 & 1+(-2) \\ 3+15 & -4+12 \\ 0+(-9) & 7+21 \end{bmatrix} = \begin{bmatrix} 8 & -1 \\ 18 & 8 \\ -9 & 28 \end{bmatrix}$$

We next define the scalar multiple kx, for a number k and a matrix x. We just multiply every entry of x by k. So if

$$x = [x_{ij}]_{1 \leq i \leq m, 1 \leq j \leq n}$$

is any $m \times n$ matrix and k is any real number then kx is another $m \times n$ matrix. Specifically

$$kx = [kx_{ij}]_{1 \leq i \leq m, 1 \leq j \leq n}$$

For example For example

$$8 \begin{bmatrix} 2 & 1 \\ 3 & -4 \\ 0 & 7 \end{bmatrix} = \begin{bmatrix} 8(2) & 8(1) \\ 8(3) & 8(-4) \\ 8(0) & 8(7) \end{bmatrix} = \begin{bmatrix} 16 & 8 \\ 24 & -32 \\ 0 & 56 \end{bmatrix}$$

We see that if we multiply by $k = 0$ we get a matrix where all the entries are 0. This has a special name.

The $m \times n$ matrix where every entry is 0 is called the $m \times n$ *zero matrix*. Thus we have zero matrices of every possible size.

If x is a matrix then we can say

$$x + \mathbf{0} = x$$

if $\mathbf{0}$ means the zero matrix of the same size as x. If we wanted to make the notation less ambiguous, we could write something like $\mathbf{0}_{m,n}$ for the $m \times n$ zero matrix. Then the things we can say is that if x is any $m \times n$ matrix then

$$x + \mathbf{0}_{m,n} = x, \qquad 0x = \mathbf{0}_{m,n}$$

We will not usually go to the lengths of indicating the size of the zero matrix we mean in this way. We will write the zero matrix as $\mathbf{0}$ and try to make it clear what size matrices we are dealing with from the context.

5.4 Matrix multiplication

This is a rather new thing, compared to the ideas we have discussed up to now. Certain matrices can be multiplied and their product is another matrix.

If x is an $m \times n$ matrix and y is an $n \times p$ matrix then the product xy will make sense and it will be an $m \times p$ matrix.

For example, then

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

is going to make sense. It is the product of

$$2 \times 3 \text{ by } 3 \times 4$$

and the result is going to be 2×4. (We have to have the same number of columns in the left matrix as rows in the right matrix. The outer numbers, the ones left after 'cancelling' the same number that occurs in the middle, give the size of the product matrix.)

Here is an example of a product that **will not be defined** and will not make sense

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 7 & 8 \\ 9 & 10 \end{bmatrix} \qquad 2 \times 3 \text{ by } 2 \times 2$$

Back to the example that will make sense, what we have explained so far is the shape of the product

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 & -2 \\ 2 & -1 & 3 & 1 \\ 4 & 2 & 6 & 4 \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & z_{13} & z_{14} \\ z_{21} & z_{22} & z_{23} & z_{24} \end{bmatrix}$$

and we still have to explain how to calculate the z_{ij}, the entries in the product. We'll concentrate on one example to try and show the idea. Say we look at the entry z_{23}, the $(2,3)$ entry in the product. What we do is take row 2 of the left matrix 'times' column 3 of the right matrix

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 & -2 \\ 2 & -1 & 3 & 1 \\ 4 & 2 & 6 & 4 \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & z_{13} & z_{14} \\ z_{21} & z_{22} & z_{23} & z_{24} \end{bmatrix}$$

The way we multiply the row $\begin{bmatrix} 4 & 5 & 6 \end{bmatrix}$ times the column

$$\begin{bmatrix} 1 \\ 3 \\ 6 \end{bmatrix}$$

is a very much reminiscent of a dot product

$$(4)(1) + (5)(3) + (6)(6) = z_{23}$$

In other words $z_{23} = 55$

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 & -2 \\ 2 & -1 & 3 & 1 \\ 4 & 2 & 6 & 4 \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & z_{13} & z_{14} \\ z_{21} & z_{22} & 55 & z_{24} \end{bmatrix}$$

If we calculate all the other entries in the same sort of way (row i on the left times colum j on the right gives z_{ij} we get

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 & -2 \\ 2 & -1 & 3 & 1 \\ 4 & 2 & 6 & 4 \end{bmatrix} = \begin{bmatrix} 17 & 4 & 25 & 12 \\ 38 & 7 & 55 & 21 \end{bmatrix}$$

The only way to get used to the way to multiply matrices is to do some practice. It is possible to explain in a succinct formula what the rule is for calculating the entries of the product matrix. In

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & & \ddots & \vdots \\ z_{m1} & z_{m2} & \cdots & z_{mp} \end{bmatrix}$$

the (i, k) entry z_{ik} of the product is got by taking the dot product of the i^{th} row $[x_{i1}\, x_{i2} \ldots x_{in}]$ of the first matrix times the k^{th} column $\begin{bmatrix} y_{1k} \\ y_{2k} \\ \vdots \\ y_{nk} \end{bmatrix}$ of the second. In short

$$x_{i1}y_{1k} + x_{i2}y_{2k} + \cdots + x_{in}y_{nk} = z_{ik}$$

If you are familiar with the Sigma notation for sums, you can rewrite this as

$$\sum_{j=1}^{n} x_{ij}y_{jk} = z_{ik} \quad (\text{for } 1 \leq i \leq m, 1 \leq k \leq p).$$

5.5 Remarks about mathematica

This might be a good time to recall that Mathematica knows how to manipulate matrices. See section 4.18.

Mathematica treats matrices using the idea of a list. Lists in Mathematica are given by curly brackets (or braces) and commas to separate the items in the list.

Mathematica uses this to indicate n-tuples of numbers (vectors in $\mathbb{R}^n$). So `p = {4, 5, 3}` would be the way to tell Mathematica you want to start talking about a point in $\mathbb{R}^3$ with coordinates $(4, 5, 3)$ or the vector $4\mathbf{i} + 5\mathbf{j} + 3\mathbf{k}$.

Mathematica understands matrices as lists of rows. So top get Mathematica to deal with

$$x = \begin{bmatrix} 3 & 4 \\ 5 & -6 \\ 7 & 8 \end{bmatrix}$$

we should instruct it to put x equal to

```
{ {3, 4}, {5, -6}, {7,8}}
```

The idea is that Mathematica views the 3×2 matrix as a list of 3 rows, and each row as a list of two numbers.

There is a fancier way to input matrices into Mathematica, so that they look like matrices as you enter them. However, Mathematica will sooner or later show you this list form. If you want to see a matrix laid out nicely, say the result of a calculation, you have to use the `MatrixForm[ ]` command on the matrix.

Adding matrices in Mathematica is easy (just use the the ordinary plus sign) and so is multiplication of matrices by scalars. However, matrix multiplication has to be done with a dot.

```
In[1]:= a = {{1, 2}, {3, 4}}

Out[1]= {{1, 2}, {3, 4}}

In[2]:= b = {{3, 5}, {1, -1}}

Out[2]= {{3, 5}, {1, -1}}

In[3]:= MatrixForm[a]

Out[3]//MatrixForm= 1   2

                    3   4

In[4]:= a+b

Out[4]= {{4, 7}, {4, 3}}

In[5]:= 2a

Out[5]= {{2, 4}, {6, 8}}

In[6]:= a b

Out[6]= {{3, 10}, {3, -4}}
```

```
In[7]:= MatrixForm[a b]

Out[7]//MatrixForm= 3    10

                    3    -4
```

This shows that Mathematica will happily give the wrong answer. Well, not wrong exactly, but not the right way to multiply matrices. Now we do it right, using the dot.

```
In[8]:= a.b

Out[8]= {{5, 3}, {13, 11}}

In[9]:= b.a

Out[9]= {{18, 26}, {-2, -2}}
```

This is an imporatant feature of matrix multiplication: ab and ba are usually different when a and b are matrices. **The order is important!**

5.6 Properties of matrix multiplication

Matrix multiplication has properties that you would expect of any multiplication. The standard rules of algebra work out, or most of them, as long as you keep the order of the products intact.

(i) If a and b are both $m \times n$ matrices and c is $n \times p$, then

$$(a + b)c = ac + bc$$

and

$$(ka)c = k(ac) = a(kc)$$

(ii) If a is an $m \times n$ matrices and b and c are both $n \times p$ and k is a scalar, then

$$a(b + c) = ab + ac$$

(iii) If a is an $m \times n$ matrices and b is $n \times p$ and c is $p \times q$, then the two ways of calculating abc work out the same:

$$(ab)c = a(bc)$$

(This is known as the associative law for multiplication.)

In the Mathematica transcript above, you see that $ab \neq ba$ in general for matrices. The situation is as follows.

(a) ba does not have to make sense if ab makes sense.

For example if a is a 3×4 matrix and b is 4×2, then ab does make sense. ab is $(3\times 4)(4\times 2)$ and so makes sense as a 3×2 matrix. But ba whould be a product of a 4×2 times a 3×4 — so it makes no sense.

(b) It can be that ab and ba both make sense but they are different sizes. For example of a is a 2×3 matrix and b is a 3×2 matrix, then ab is 2×2 while ba is 3×3. As they are different sizes ab and ba are certainly not equal.

(c) The more tricky case is the case where the matrices a and b are *square matrices* of the same size.

A square matrix is an $n\times n$ matrix for some n. Notice that the product of two $n\times n$ matrices is another $n\times n$ matrix.

Still, it is usually not the case that $ab=ba$ when a and b are $n\times n$. The example we worked out with Mathematica was

$$a=\begin{bmatrix}1&2\\3&4\end{bmatrix}, b=\begin{bmatrix}3&5\\1&-1\end{bmatrix}, ab=\begin{bmatrix}5&3\\13&11\end{bmatrix}, ba=\begin{bmatrix}18&26\\-2&-2\end{bmatrix}$$

The upshot is that **the order matters** in matrix multiplication. The last example is not at all hard to come up with. If you write down two $n\times n$ matrices a and b at random, the chances are $ab\neq ba$.

There are some special square matrices which deserve a special name. We've already seen the zero matrix (which makes sense for any size — can be $m\times n$ and need not be square). One special matrix is the $n\times n$ *identity matrix* which we denote by I_n. So

$$I_2=\begin{bmatrix}1&0\\0&1\end{bmatrix}, I_3=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}, I_4=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix}$$

and in general I_n is the $n\times n$ matrix with 1 in all the 'diagonal' entries and zeroes off the diagonal.

By the *diagonal entries* of an $n\times n$ matrix we mean the (i,i) entries for $i=1,2,\dots,n$. We try not to talk of the diagonal for rectangular matrices (because the line from the top left corner to the bottom right probably won't contain many entries of the matrix).

The reason for the name is that the identity matrix is a multiplicative identity. That is $I_m a=a$ and $a=aI_n$ for any $m\times n$ matrix a. These facts are easy to figure out.

5.7 Systems of linear equations revisited

There is a way to write a system of linear equations as a single matrix equation. For example, the system

$$\begin{array}{rcrcrcrcr}5x_1&-&2x_2&+&x_3&-&4x_4&=&-3\\2x_1&+&3x_2&+&7x_3&+&2x_4&=&18\\x_1&+&2x_2&-&x_3&-&x_4&=&-3\end{array}$$

of 3 equations in 4 unknowns can be written

$$\begin{bmatrix} 5x_1 - 2x_2 + x_3 - 4x_4 \\ 2x_1 + 3x_2 + 7x_3 + 2x_4 \\ x_1 + 2x_2 - x_3 - x_4 \end{bmatrix} = \begin{bmatrix} -3 \\ 18 \\ -3 \end{bmatrix}$$

and the left side can be written as a matrix product. We get

$$\begin{bmatrix} 5 & -2 & 1 & -4 \\ 2 & +3 & +7 & +2 \\ 1 & +2 & -1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -3 \\ 18 \\ -3 \end{bmatrix}$$

This has the form

$$A\mathbf{x} = \mathbf{b} \tag{1}$$

where

$$A = \begin{bmatrix} 5 & -2 & 1 & -4 \\ 2 & 3 & 7 & 2 \\ 1 & 2 & -1 & -1 \end{bmatrix}$$

is the matrix of the coefficients for the unknowns x_1, x_2, x_3, x_4 (a 3×4 matrix),

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

is the 4×1 (column) matrix made up of the unknowns, and

$$\mathbf{b} = \begin{bmatrix} -3 \\ 18 \\ -3 \end{bmatrix}$$

is the (3×1 column) matrix of the constant terms (right hand sides) in the system of linear equations.

Note that this is rather different from the augmented-matrix shorthand we used in Chapter 1. That could be summarised as taking the matrix

$$[A \,|\, \mathbf{b}],$$

which is a $3 \times (4 + 1) = 3 \times 5$ matrix

$$A = \begin{bmatrix} 5 & -2 & 1 & -4 & : -3 \\ 2 & 3 & 7 & 2 & : \ 18 \\ 1 & 2 & -1 & -1 & : -3 \end{bmatrix}$$

Recall that the positions of the entries in the augmented matrix corresponds to the rôle of the number as a coefficent in the system of equations, while the dotted line is there to remind us of the position of the equals sign.

Looking at the equation (1), you should be remined of the simplest possible linear equations in a single unknowm, like $5x = 21$, which we solve by dividing across by the thing multiplying x. (In the example $5x = 21$ we divide across by 5, or mulriuply both sides of the equation by $\frac{1}{5}$ to get the solution $x = \frac{21}{5}$.)

Thinking in these terms, it seems tempting to solve the equation (1) by 'dividing' both sides by A. One problem is to make sense of division by a matrix. That would be the same as making sense of the reciprocal of the matrix, or one over the matrix.

In the actual example we picked, with fewer equations than unknowns, this idea is never going to work. We know from before that when we simplify 3 equations in 4 unknowns via Gauss-Jordan elimination, one of two things can happen. Either we have inconsistent equations (with no solutions at all) or we will end up with at least 1 free variable.

However, if we did have the same number of equations as unknowns, we are quite often going to end up with just one solution for the unknowns. That is the case where we can possibly have a reciprocal for the matrix A that comes up, except we will call it the inverse rather than the reciprocal.

To summarise the point here, it is that a system of m linear equations in n unknowns

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \vdots \qquad\qquad\qquad\qquad\qquad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \end{cases}$$

can be written as a single matrix equation

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & \ddots & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

So it is of the form (1) where now A is an $m \times n$ matrix, $\mathbf{x}$ is an $n \times 1$ (column) matrix of unknowns and $\mathbf{b}$ is an $m \times 1$ column.

5.8 Inverse matrces — basic ideas

Definition: If A is an $n \times n$ matrix, then another $n \times n$ matrix C is called the *inverse matrix* for A if it satisfies

$$AC = I_n \text{ and } CA = I_n.$$

We write A^{-1} for the inverse matrix C (if there is one).

The idea for this definition is that the identity matrix is analagous to the number 1, in the sense that $1k = k1 = k$ for every real number k while $AI_n = I_nA = A$ for every $n \times n$ matrix

A. (That's why it is called the identity matrix.) Then the key thing about the reciprocal of a nonzero number k is that the product

$$\left(\frac{1}{k}\right) k = 1$$

For numbers the order of the product is not important, but for matrices the order matters. That is why we insist that the inverse should work on both sides.

A bit later on though, we will see a theorem that says that if A and C are $n \times n$ matrices and $AC = I_n$, then automatically $CA = I_n$ must also hold. Because AC is usually not the same as CA, it should not be expected that $AC = CA$ when $AC = I_n$. But it is true (for square matrices).

However, one thing that is not as obvious as for numbers, is when there is an inverse for a given matrix A. It is not enough that A should be nonzero. One way to see this is to look at a system of n linear equations in n unknowns written in the matrix form (1). If the $n \times n$ matrix A has an inverse matrix C then we can multiply both sides of the equation (1) by C from the left to get

$$\begin{aligned} A\mathbf{x} &= \mathbf{b} \\ C(A\mathbf{x}) &= C\mathbf{b} \\ (CA)\mathbf{x} &= C\mathbf{b} \\ I_n\mathbf{x} &= C\mathbf{b} \\ \mathbf{x} &= C\mathbf{b} \end{aligned}$$

So we find that the system of n equation in n unknowns given by (1) will just have the one solution $\mathbf{x} = C\mathbf{b}$. And that will be true for any right hand side $\mathbf{b}$.

This reveals a special property for an $n \times n$ matrix A. It means that there are really n equations in (1), none are dependent on the others, none inconsistent with the others. This amounts to a significant restriction on A.

Definition. An $n \times n$ matrix A is called **invertible** if there is an $n \times n$ inverse matrix for A.

5.9 Finding inverse matrices

We now consider how to find the inverse of a given matrix A. The method we explain will work quite efficiently for large matrices as well as for small ones.

We'll leave aside the question of whether there is an inverse for the square matrix A that we start with. We will also just look for C by looking at the equation $AC = I_n$, and worry later about the claim we made before that $CA = I_n$ will work out automatically once $AC = I_n$.

To make things more concrete, we'll thing about a specific exaample

$$A = \begin{bmatrix} 2 & 3 \\ 2 & 5 \end{bmatrix}$$

We think of how we can find

$$C = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$$

so that $AC = I_2$. Writing out that equation we want C to satify we get

$$AC = \begin{bmatrix} 2 & 3 \\ 2 & 5 \end{bmatrix} \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I_2$$

If you think of how matrix multiplication works, this amounts to two different equations for the columns of C

$$\begin{bmatrix} 2 & 3 \\ 2 & 5 \end{bmatrix} \begin{bmatrix} c_{11} \\ c_{21} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \text{ and } \begin{bmatrix} 2 & 3 \\ 2 & 5 \end{bmatrix} \begin{bmatrix} c_{12} \\ c_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

According to the reasoning we used above to get to equation (1), each of these represents a system of 2 linear equations in 2 unknowns that we can solve for the unknowns, and the unknonws in this case are the columns of C.

We know then how to solve them. We can use Gauss-Jordan elimination (or Gaussian elimination) twice, once for the augmented matrix for the first system of equations,

$$\left[\begin{array}{cc:c} 2 & 3 & 1 \\ 2 & 5 & 0 \end{array}\right]$$

and again for the second system

$$\left[\begin{array}{cc:c} 2 & 3 & 0 \\ 2 & 5 & 1 \end{array}\right]$$

If we were to write out the steps for the Gauss-Jordan eliminations, we'd find that we were repeating the exact same steps the second time as the first time. The same steps, but the column to the right of the dotted line will be different in each case. There is a trick to solve at once two systems of linear equations, where the coefficients of the unknowns are the same in both, but the right hand sides are different. (That is the situation we have.) The trick is to write both columns after the dotted line, like this

$$\left[\begin{array}{cc:cc} 2 & 3 & 1 & 0 \\ 2 & 5 & 0 & 1 \end{array}\right]$$

We row reduce this matrix

$$\left[\begin{array}{cc:cc} 1 & \frac{3}{2} & \frac{1}{2} & 0 \\ 2 & 5 & 0 & 1 \end{array}\right] \text{OldRow1} \times \tfrac{1}{2}$$

$$\left[\begin{array}{cc:cc} 1 & \frac{3}{2} & \frac{1}{2} & 0 \\ 0 & 2 & -1 & 1 \end{array}\right] \text{OldRow2} - 2 \times \text{OldRow1}$$

$$\left[\begin{array}{cc:cc} 1 & \frac{3}{2} & \frac{1}{2} & 0 \\ 0 & 1 & -\frac{1}{2} & \frac{1}{2} \end{array}\right] \text{OldRow2} \times \tfrac{1}{2}$$

(row echolon form now)

$$\left[\begin{array}{cc:cc} 1 & 0 & \frac{5}{4} & -\frac{3}{4} \\ 0 & 1 & -\frac{1}{2} & \frac{1}{2} \end{array}\right] \text{OldRow1} - \tfrac{3}{2} \times \text{OldRow2}$$

This is in reduced row echelon form. (Gauss-Jordan finished.)

The first column after the dotted line gives the solution to the first system, the one for the first column of C. The second column after the dotted line relates to the second system, the one for the second column of C. That means we have

$$\begin{bmatrix} c_{11} \\ c_{21} \end{bmatrix} = \begin{bmatrix} \frac{5}{4} \\ -\frac{1}{2} \end{bmatrix} \text{ and } \begin{bmatrix} c_{12} \\ c_{22} \end{bmatrix} = \begin{bmatrix} -\frac{3}{4} \\ \frac{1}{2} \end{bmatrix}$$

So we find that the matrix C has to be

$$C = \begin{bmatrix} \frac{5}{4} & -\frac{3}{4} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

We can multiply out and check that it is indeed true that $AC = I_2$ (which has to be the case unless we made a mistake) and that $CA = I_2$ (which has to be true automatically according to a theorem that we have mentioned is coming later).

$$AC = \begin{bmatrix} 2 & 3 \\ 2 & 5 \end{bmatrix} \begin{bmatrix} \frac{5}{4} & -\frac{3}{4} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 2\left(\frac{5}{4}\right) + 3\left(-\frac{1}{2}\right) & 2\left(-\frac{3}{4}\right) + 3\left(\frac{1}{2}\right) \\ 2\left(\frac{5}{4}\right) + 5\left(-\frac{1}{2}\right) & 2\left(-\frac{3}{4}\right) + 5\left(\frac{1}{2}\right) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$CA = \begin{bmatrix} \frac{5}{4} & -\frac{3}{4} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 2 & 5 \end{bmatrix} = \begin{bmatrix} \left(\frac{5}{4}\right)(2) + \left(-\frac{3}{4}\right)(2) & \left(\frac{5}{4}\right)(3) + \left(-\frac{3}{4}\right)(5) \\ \left(-\frac{1}{2}\right)(2) + \left(\frac{1}{2}\right)(2) & \left(-\frac{1}{2}\right)(3) + \left(\frac{1}{2}\right)(5) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

As mentioned before, this approach works for larger matrices too. If we start with an $n \times n$ matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

and we look for an $n \times n$ matrix

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix}$$

where $AC = I_n$, we want

$$AC = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = I_n$$

This means that the columns of C have to satisfy systems of n linear equations in n unknowns of the form

$$A(j^{\text{th}} \text{ column of } C) = j^{\text{th}} \text{ column of } I_n$$

We can solve all of these n systems of equations together becuase they have the same matrix A of coefficients for the unknowns. We do this by writing an augmented matrix where there are n columns after the dotted line. The columns to the right of the dotted line, the right hand sides of the various systems we want to solve to find the columns of C are going to be the columns of the $n \times n$ identity matrix. Summarising, this is what we have.

Method: (Method for finding the inverse A^{-1} of an $n \times n$ matrix A.) Use Gauss-Jordan elimination to row reduce the augmented matrix

$$[A \,|\, I_n] = \left[\begin{array}{cccc:cccc} a_{11} & a_{12} & \cdots & a_{1n} & : & 1 & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & a_{2n} & : & 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots & : & \vdots & & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & : & 0 & 0 & \cdots & 1 \end{array}\right]$$

We should end up with a reduced row echelon form that looks like

$$\left[\begin{array}{cccc:cccc} 1 & 0 & \cdots & 0 & : & c_{11} & c_{12} & \cdots & c_{1n} \\ 0 & 1 & \cdots & 0 & : & c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & & \ddots & \vdots & : & \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & : & c_{n1} & c_{n2} & \cdots & c_{nn} \end{array}\right]$$

or in summary $[I_n \,|\, A^{-1}]$.

We'll now look into when this works more carefully. If we don't end up with a matrix of the form $[I_n \,|\, C]$ it means that there is no inverse for A.

5.10 Elementary matrices

We now make a link between elementary row operations and matrix multiplication. Recall now the 3 types of elementary row operations as laid out in section 1.6.

(i) multiply all the numbers is some row by a nonzero factor (and leave every other row unchanged)

(ii) replace any chosen row by the difference between it and a multiple of some other row.

(iii) Exchange the positions of some pair of rows in the matrix.

Definition: An $n \times n$ *elementary matrix* E is the result of applying a single elementary row operation to the $n \times n$ identity matrix I_n.

Examples. We use $n = 3$ in these examples. Recall

$$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

(i) Row operation: Multiply row 2 by -5. Corresponding elementary matrix

$$E = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -5 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

(ii) Row operation: Add 4 times row 1 to row 3 (same as subtracting (-4) times row 1 from row 3). Corresponding elementary matrix

$$E = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 4 & 0 & 1 \end{bmatrix}$$

(iii) Row operation: swop rows 2 and 3.

$$E = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

5.11 Link of matrix multiplication to row oerations

The idea here is that if A is an $m \times n$ matrix, then doing one single row operation on A is equivalent to multiplying A on the left by an elementary matrix E (to get EA), and E should be the $m \times m$ elementary matrix for that same row operation.

Examples. We use the following A to illustrate this idea,

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{bmatrix}$$

(i) Row operation: Add (-5) times row 1 to row 2. Corresponding EA is

$$EA = \begin{bmatrix} 1 & 0 & 0 \\ -5 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & -4 & -8 & -12 \\ 9 & 10 & 11 & 12 \end{bmatrix}$$

(Same as doing the row operation to A.)

(ii) Row operation: Suppose in addition we also want to add (-9) times row 1 to row 3. We've been doing two steps together, but really they should be done one at a time. (Doing two together is ok as long as it is clear that you could still do the second one after you've done the first.) In the context of multiplying by elementary matrices, we need a different elementary matrix for the second step

$$E_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -9 & 0 & 1 \end{bmatrix}$$

What we want in order to do first one and then the next row operation is

$$E_2EA = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -9 & 0 & 1 \end{bmatrix} EA = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -9 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & -4 & -8 & -12 \\ 9 & 10 & 11 & 12 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & -4 & -8 & -12 \\ 0 & -8 & -16 & -24 \end{bmatrix}$$

where E is the elementary matrix we used first.

There is a justification for going back and renaming the first one E_1 rather than E. So the first row operation changes A to E_1A, and then the second changes that to E_2E_1A.

If we do a whole sequence of several row operations (as we would do if we followed the Gaussian elimination recipe further) we can say that the end result after k row operations is that we get

$$E_kE_{k-1}\dots E_3E_2E_1A$$

where E_i is the elementary matrix for the i^{th} row operation we did.

5.12 Elementary matrices are invertible

As we explained at the end of section 1.5, all elementary row operations are reversible by another elementary row operation. It follows that every elementray matrix E has an inverse that is another elementary matrix.

For example, take E to be the 3×3 elementary matrix corresonding the the row operation "add (-5) times row 1 to row 2". So

$$E = \begin{bmatrix} 1 & 0 & 0 \\ -5 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Then the reverse row operation is "add 5 times row 1 to row 2", and the elementary matrix for that is

$$\tilde{E} = \begin{bmatrix} 1 & 0 & 0 \\ 5 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Thinking in terms of row operations, or just my multiplying out the matrices we see that

$$\tilde{E}E = \text{ result of applying second row operation to } E = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = I_3$$

and $E\tilde{E} = I_3$ also.

5.13 Theory about invertible matrices

5.13.1 Theorem. *Products of invertible matrices are invertible, and the inverse of the product is the product of the inverses taken in the reverse order.*

In more mathematical language, if A and B are two invertible $n \times n$ matrices, then AB is invertible and $(AB)^{-1} = B^{-1}A^{-1}$.

Proof. Start with any two invertible $n \times n$ matrices A and B, and look at

$$(AB)(B^{-1}A^{-1}) = A(BB^{-1})A^{-1} = AI_nA^{-1} = AA^{-1} = I_n$$

And look also at

$$(B^{-1}A^{-1})(AB) = B^{-1}(B^{-1}B)A = B^{-1}I_nB = B^{-1}B = I_n$$

This shows that $B^{-1}A^{-1}$ is the inverse of AB (because multiplying AB by $B^{-1}A^{-1}$ on the left or the right gives I_n). So it shows that $(AB)^{-1}$ exists, or in other words that AB is invertible, as well as showing the formula for $(AB)^{-1}$. □

5.13.2 Theorem (ways to see that a matrix is invertible)**.** *Let A be an $n \times n$ (square) matrix.*

The following are euivalent statemnts about A, meaning that is any one of them is true, then the other have to be true as well. (And if one is not true, the others must all be not true.)

(a) A is invertible (has an inverse)

(b) the equation $A\mathbf{x} = \mathbf{0}$ (where $\mathbf{x}$ is an umknown $n \times 1$ column matrix, $\mathbf{0}$ is the $n \times 1$ zero column) has only the solution $\mathbf{x} = \mathbf{0}$

(c) the reduced row echelon for for A is I_n

(d) A can be written as a product of elementary matrices

Proof. We'll actually relegate the proof to an appendix, even though we are now in a position to explain the reasons the theorem works. The details seemed a bit lengthy and abstract to go through them in the lectures, even though they just involve putting together things we have already done, and the book by Anton & Rorres goes through this proof.

One thing that we will explain here is the overall way of approaching the proof.

The whole idea of what a proof is in Mathematics should be borne in mind. Theorems are the mathematical version of the laws of science (the second law of thermodynamics, Boyle's law, Newtons Laws and so on), but there is a difference. In Science, somebody formulates a possible rule or law as a way of summarising observations made in experiments. The law should then be checked with further experiments and if it checks out, it becomes accepted as a fact. Such laws generally have to have a precise statement for them to work. Roughly they say that given a certain situation, some particular effect or result will happen. Sometimes the "certain situation" may be somewhat idealised. For example, some things may hold in a vaccuum, or in the absence of gravity, and these circumstances are hard to come by in a perfect sense. So one may interpret

the law as saying that the effect or result should be very close to the obererved effect or result if the situation is almost exactly valid. So it is true to say that light travels in a straight line (in a vacuum and in the absence of gravity), and it is almost true even if there is gravity. But over long distances across space, light can be observed to have been bent.

In mathematics we expect our theorems to be exactly true as stated. So there will be assumptions about the situation (certain kind of matrix, certain kind of function, maybe a combination of several assumptions). But then the idea is that the conclusion of the theorem should *always* hold when the assumptions are valid. We don't check a theorem by experience, or by experiments. We might realise it is possibly true on such a basis, but the idea then is to show by some steps of logical reasoning that the conclusion must always hold in the situation where the assumptions are valid.

In principle there are a lot of things to prove in the theorem we are dicussing. Staring with any one of the 4 items, assuming that that statement is valid for a given $n \times n$ matrix A, we should provide a line of logical reasoning why all the other items have to be also true about that same A. We don't do this by picking examples of matrices A, but by arguing about a matrix where we don't specifically know any of the entries. But we then have 4 times 3 little proofs to give, 12 proofs in all. So it would be long even if each individual proof is very easy.

There is a trick to reduce the number of proofs from 12 to only 4. We prove a cyclical number of steps

$$\begin{array}{ccc} \text{(a)} & \Rightarrow & \text{(b)} \\ \Uparrow & & \Downarrow \\ \text{(d)} & \Leftarrow & \text{(c)} \end{array}$$

The idea then is to prove 4 things only

(a) $\Rightarrow$ **(b)** In this step we assume only that statement (a) is true about A, and then we show that (b) must also be true.

(b) $\Rightarrow$ **(c)** In this step we assume only that statement (b) is true about A, and then we show that (c) must also be true.

(c) $\Rightarrow$ **(d)** Similarly we assume (c) and show (d) must follow.

(d) $\Rightarrow$ **(a)** In the last step we assume (d) (not any of the others, only (d)) and show that (a) must follow.

When we have done this we will be able to deduce all the statements from any one of the 4. Starting with (say) the knowledge that (c) is a true statement the third step above shows that (d) must be true. Then the next step tells us (a) must be true and the first step then says (b) must be true. In other words, starting at any point around the ring (or at any corner of the square) we can work around to all the others.

We'll leave the 4 proofs out though, but give them in an appendix in case you are intereset. □

5.13.3 Theorem. *If A and B are two $n \times n$ matrices and if $AB = I_n$, then $BA = I_n$.*

Proof. The idea is to apply Theorem 5.13.2 to the matrix B rather than to A.

Consider the equation $B\mathbf{x} = \mathbf{0}$ (where $\mathbf{x}$ and $\mathbf{0}$ are $n \times 1$). Multiply that eqquation by A on the left to get

$$\begin{aligned} AB\mathbf{x} &= B\mathbf{0} \\ I_n\mathbf{x} &= \mathbf{0} \\ \mathbf{x} &= \mathbf{0} \end{aligned}$$

So $\mathbf{x} = \mathbf{0}$ is the only possible solution of $B\mathbf{x} = \mathbf{0}$.

That means B must satisfy condition (b) of Theorem 5.13.2. Thus by the theorem, B^{-1} exists. Multiply the equation $AB = I_n$ by B^{-1} on the right to get

$$\begin{aligned} ABB^{-1} &= I_nB^{-1} \\ AI_n &= B^{-1} \\ A &= B^{-1} \end{aligned}$$

So, we get

$$BA = BB^{-1} = I_n.$$

□

5.14 Special matrices

There are matrices that have a special form that makes calculations with them much easier than the same calculations are as a rule.

Diagonal matrices For square matrices (that is $n \times n$ for some n) $A = (a_{ij})_{i,j=1}^n$ we say that A is a *diagnonal matrix* if $a_{ij} = 0$ whenever $i \neq j$. Thus in the first few cases $n = 2, 3, 4$ diagonal matrices look like

$$\begin{bmatrix} a_{11} & 0 \\ 0 & a_{22} \end{bmatrix}$$

$$\begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix}$$

$$\begin{bmatrix} a_{11} & 0 & 0 & 0 \\ 0 & a_{22} & 0 & 0 \\ 0 & 0 & a_{33} & 0 \\ 0 & 0 & 0 & a_{44} \end{bmatrix}$$

Examples with numbers

$$\begin{bmatrix} 4 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 13 \end{bmatrix}, \qquad \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 4 \end{bmatrix}.$$

(These are 3×3 examples.)

Diagonal matrices are easy to multiply

$$\begin{bmatrix} 4 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 6 \end{bmatrix} \begin{bmatrix} -1 & 0 & 0 \\ 0 & 12 & 0 \\ 0 & 0 & 4 \end{bmatrix} = \begin{bmatrix} -4 & 0 & 0 \\ 0 & 60 & 0 \\ 0 & 0 & 24 \end{bmatrix}$$

$$\begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix} \begin{bmatrix} b_{11} & 0 & 0 \\ 0 & b_{22} & 0 \\ 0 & 0 & b_{33} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & 0 & 0 \\ 0 & a_{22}b_{22} & 0 \\ 0 & 0 & a_{33}b_{33} \end{bmatrix}$$

The idea is that all that needs to be done is to multiply the corresponding diagonal entries to get the diagonal entries of the product (which is again diagonal).

Based on this we can rather easily figure out how to get the inverse of a diagonal matrix. For example if

$$A = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 6 \end{bmatrix}$$

then

$$A^{-1} = \begin{bmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{5} & 0 \\ 0 & 0 & \frac{1}{6} \end{bmatrix}$$

because if we multiply these two diagonal matrices we get the identity.

We could also figure out A^{-1} the usual way, by row-reducing $[A \mid I_3]$. The calculation is actually quite easy. Starting with

$$[A \mid I_3] = \left[\begin{array}{ccc:ccc} 4 & 0 & 01 & : & 0 & 0 & \\ 0 & 5 & 00 & : & 1 & 0 & \\ 0 & 0 & 6 & : & 0 & 0 & 1 \end{array}\right]$$

we just need to divide each row by something to get to

$$[A \mid I_3] = \left[\begin{array}{ccc:ccc} 1 & 0 & 0 & : & \frac{1}{4} & 0 & 0 \\ 0 & 1 & 0 & : & 0 & \frac{1}{5} & 0 \\ 0 & 0 & 1 & : & 0 & 0 & \frac{1}{6} \end{array}\right]$$

In summary, for 3×3 diagonal matrices,

$$\begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{a_{11}} & 0 & 0 \\ 0 & \frac{1}{a_{22}} & 0 \\ 0 & 0 & \frac{1}{a_{33}} \end{bmatrix}$$

and the digonal matrices that are invertible are those for which this formula makes sense — in other words, those where the diagonal entries are all non-zero, or

$$a_{11}a_{22}a_{33} \neq 0$$

A similar result holds for 2×2 diagonal matrices and for diagonal matrices of larger sizes. The number which must be non-zero for a diagonal matrix to be invertible, the product of the diagonal entries for a diagonal matrix, is an example of a "determinant". We will come to determinants (for all square matrices) in the next chapter.

Upper triangular matrices This is the name given to square matrices where all the non-zero entries are on or above the diagonal.

A 4×4 example is

$$A = \begin{bmatrix} 4 & -3 & 5 & 6 \\ 0 & 3 & 7 & -9 \\ 0 & 0 & 0 & 6 \\ 0 & 0 & 0 & -11 \end{bmatrix}$$

Another way to express it is that all the entries that are definitely below the diagonal have to be 0. Some of those on above the diagonal can be zero also. They can all be zero and then we would have the zero matrix, which would be technically upper triangular. All diagonal matrices are also counted as upper triangular.

The precise statement then is that an $n \times n$ matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

is *upper triangular* when

$$a_{ij} = 0 \text{ whenever } i > j.$$

It is fairly easy to see that if A and B are two $n \times n$ upper triangular matrices, then

$$\text{the sum } A + B \text{ and the product } AB$$

are both upper triangular.

Also inverting upper traingular matrices is relatively painless because the Gaussian elimination parts of the process are almost automatic. As an example, we look at the (upper triangular)

$$A = \begin{bmatrix} 3 & 4 & 5 & 6 \\ 0 & 7 & 8 & 9 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

We should row reduce

$$\left[\begin{array}{cccc:cccc} 3 & 4 & 5 & 6 & 1 & 0 & 0 & 0 \\ 0 & 7 & 8 & 9 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 2 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 & 0 & 1 \end{array}\right]$$

and the first few steps are to divide row 1 by 3, row 2 by 7 and row 4 by 3, to get

$$\left[\begin{array}{cccc:cccc} 1 & \frac{4}{3} & \frac{5}{3} & 2 & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 1 & \frac{8}{7} & \frac{9}{7} & 0 & \frac{1}{7} & 0 & 0 \\ 0 & 0 & 1 & 2 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & \frac{1}{3} \end{array}\right]$$

This is then already in row echelon form and to get the inverse we need to get to reduced row echelon form (starting by clearing out above the last leading 1, then working back up). The end result should be

$$\left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & \frac{1}{3} & -\frac{4}{21} & -\frac{1}{7} & 0 \\ 0 & 1 & 0 & 0 & 0 & \frac{1}{7} & -\frac{8}{7} & \frac{1}{3} \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & -\frac{2}{3} \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & \frac{1}{3} \end{array}\right]$$

It is quite easy to see that an upper traigular matrix is invertible exactly when the diagonal entries are all nonzero. Another way to express this same thing is that the product of the diagonal entries should be nonzero.

It is also easy enough to see from the way the above calculation of the inverse worked out that the inverse of an upper triangular matrix will be again upper triangular.

Strictly upper triangular matrices These are matrices which are upper triangular and also have all zeros on the diagonal. This can also be expressed by saying that there should be zeros on and below the diagonal.

The precise statement then is that an $n \times n$ matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

is *strictly upper triangular* when

$$a_{ij} = 0 \text{ whenever } i \geq j.$$

An example is

$$A = \begin{bmatrix} 0 & 1 & 2 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{bmatrix}$$

This matrix is certainly *not* invertible. To be invertible we need *each* diagonal entry to be nonzero. This matrix is at the other extreme in a way — all diagonal entries are 0.

For this matrix

$$A^2 = AA = \begin{bmatrix} 0 & 1 & 2 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 2 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

and

$$A^3 = AA^2 = \begin{bmatrix} 0 & 1 & 2 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \mathbf{0}$$

In fact this is not specific to the example. Every strictly upper triangular matrix

$$A = \begin{bmatrix} 0 & a_{12} & a_{13} \\ 0 & 0 & a_{23} \\ 0 & 0 & 0 \end{bmatrix}$$

has

$$A^2 = \begin{bmatrix} 0 & 0 & a_{12}a_{23} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \text{ and } A^3 = \mathbf{0}.$$

In general an $n \times n$ strictly upper triangular matrix A has $A^n = \mathbf{0}$.

5.14.1 Definition. A square matrix A is called *nilpotent* if some power of A is the zero matrix.

We have just seen that $n \times n$ strictly upper triangular matrices are nilpotent.

This shows a significant difference between ordinary multiplication of numbers and matrix multiplication. It is not true that $AB = \mathbf{0}$ means that A or B has to be $\mathbf{0}$. The question of which matrices have an inverse is also more complicated than it is for numbers. Every nonzero number has a reciprocal, but there are many nonzero matrices that fail to have an inverse.

5.15 Transposes

Now that we've discussed upper triangular matrices (and strictly upper triangular), it might cross your mind that we coudl also discuss lower triangular matrices. In fact we could repeat most of the same argumenst for them, with small modifications, but the transpose provides a way to flip from one to the other.

In summary the transpose of a matrix is what you get by writing toe rows as columns. More precisely, we can take the transpose of any $m \times n$ matrix A. If

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

we write the entries of the first row $a_{11}, a_{12}, \ldots, a_{1n}$ down the first column of the transpose, the entries $a_{21}, a_{22}, \ldots, a_{2n}$ of the second row down the second column, etc. We get a new matrix, which we denote A^t and is an $n \times m$ matrix

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

Another way to describe it is that the (i, j) entry of the tranpsose in a_{ji} = the (j, i) entry of the original matrix.

Examples are

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}, \quad A^t = \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \\ a_{13} & a_{32} \end{bmatrix}$$

$$A = \begin{bmatrix} 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{bmatrix}, \quad A^t = \begin{bmatrix} 4 & 7 & 10 \\ 5 & 8 & 11 \\ 6 & 9 & 12 \end{bmatrix}$$

Another way to describe it is that it is the matrix got by reflecting the original matrix in the "diagonal" line, or the line were $i = j$ (row number = column number).

So we see that if we start with an upper traingular

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix}$$

then the transpose

$$A = \begin{bmatrix} a_{11} & 0 & 0 \\ a_{12} & a_{22} & 0 \\ a_{13} & a_{23} & a_{33} \end{bmatrix}$$

is *lower triangular* (has all nonzero entries on or below the diagonal).

5.15.1 Facts about transposes

(i) $A^{tt} = A$ (transpose twice gives back the original matrix)

(ii) $(A + B)^t = A^t + B^t$ (if A and B are matrices of the same size).

This is pretty easy to see.

(iii) $(kA)^t = kA^t$ for A a matrix and k a scalar. (Again it is quite easy to see that this always works out.)

(iv) $(AB)^t = B^tA^t$ (the transpose of a product is the product of the transposes taken in the reverse order — provided the product AB makes sense).

So if A is $m \times n$ and B is $n \times p$, then $(AB)^t = B^tA^t$. Note that B^t is $p \times n$ and A^t is $n \times m$ so that B^tA^t makes sense and is a $p \times m$ matrix, the same size as $(AB)^t$.

The proof for this is a little more of a challenge to write out than the previous things. It requires a bit of notation and organisation to get it straight. So we won't do it in detail. Here is what we would need to do just for the 2×2 case.

Take any two 2×2 matrices, which we write out as

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$$

Then

$$A^t = \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{bmatrix}, \quad B^t = \begin{bmatrix} b_{11} & b_{21} \\ b_{12} & b_{22} \end{bmatrix}$$

and we can find

$$AB = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix}, \quad (AB)^t = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{21}b_{11} + a_{22}b_{21} \\ a_{11}b_{12} + a_{12}b_{22} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix}$$

while

$$B^tA^t = \begin{bmatrix} b_{11}a_{11} + b_{21}a_{12} & b_{11}a_{21} + b_{21}a_{22} \\ b_{12}a_{11} + b_{22}a_{12} & b_{12}a_{21} + b_{22}a_{22} \end{bmatrix} = (AB)^t$$

(v) If A is an invertible square matrix then A^t is also invertible and $(A^t)^{-1} = (A^{-1})t$ (the inverse of the tranpose is the same as the tranpose of the inverse.

Proof. Let A be an invertibe $n \times n$ matrix. We know from the definition of A^{-1} that

$$AA^{-1} = I_n \text{ and } A^{-1}A = I_n$$

Take transposes of both equations to get

$$(A^{-1})^tA^t = I_n^t = I_n \text{ and } A^t(A^{-1})^t = I_n^t = I_n$$

Therefore we have proved that A^t has an inverse and that the inverse matrix is $(A^{-1})^t$. So we have proved the formula $(A^t)^{-1} = (A^{-1})t$. □

5.16 Lower triangular matrices

We can use the transpose to transfer what we know about upper triagnular matrices to lower triangular ones. Let us take 3×3 matrices as an example, though what we say will work similarly for $n \times n$.

If

$$A = \begin{bmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

is lower traingular, then its transpose

$$A^t = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ 0 & a_{22} & a_{32} \\ 0 & 0 & a_{33} \end{bmatrix}$$

is upper triangular. So we know that A^t has an inverse exactly when the product of its diagonal entries

$$a_{11}a_{22}a_{33} \neq 0$$

But that is the same as the product of the diagonal entries of A.

So lower traingular matrices have an inverse exactly when the product of the diagonal entries is nonzero.

Another thing we know is that $(A^t)^{-1}$ is again upper triangular. So $((A^t)^{-1})^t = (A^{-1})^{tt} = A^{-1}$ is lower triangular. In this way we can show that the inverse of a lower traingualr matrix is again lower traingular (if it exists).

Using $(AB)^t = B^tA^t$ we could also show that the product of lower triangular matrices [of the same size] is again lower triangular. (The idea is that $B^tA^t =$ a product of upper triangulars is upper triangular and then $AB = ((AB)^t)^t = (B^tA^t)^t =$ transpose of upper triangular and so AB is lower triangular.) You can also figure this out by seeing what happens when you multiply two lower traingular matrices together.

Finally, we could use transpsoes to show that stricly *lower triangular matrices* have to be nilpotent (some power of them is the zero matrix). Or we could figure that out by working it out in more or less the same way as we did for the strictly upper triangular case.

5.17 Trace of a matrix

In the next chapter we will see how to work out a determinant for any square matrix A, a number that 'determines' whether or not A is invertible. For diagonal and triangular matrices (upper or lower traingular) we already have such a number, thr product of the diagonal entries. It will be more complicated to work out though when we look at more complicated matrices.

The *trace* of a matrix is a number that is quite easy to compute. It is the sum of the diagonal entries. So

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

has

$$\text{trace}(A) = 1 + 4 = 5$$

and

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ -7 & -8 & -6 \end{bmatrix}$$

has

$$\text{trace}(A) = 1 + 5 + (-6) = 0$$

For 2×2

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \Rightarrow \text{trace}(A) = a_{11} + a_{22}$$

and for 3×3,

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \Rightarrow \text{trace}(A) = a_{11} + a_{22} + a_{33}$$

Although the traceof a matrix is easy to calcualte, it is not that wonderfully useful. The properties it has are as follows.

5.17.1 Properties of the trace

(i) $\text{trace}(A + B) = \text{trace}(A) + \text{trace}(B)$ (if A and B are both $n \times n$)

(ii) $\text{trace}(kA) = k\,\text{trace}(A)$ (if k is a scalara and A is a square matrix)

(iii) $\text{trace}(A^t) = \text{trace}(A)$ (if A is any square matrix)

(iv) $\text{trace}(AB) = \text{trace}(BA)$ for A and B square matrices of the same size (or even for A $m \times n$ and B an $n \times m$ matrix).

The last property is the only one that is at all hard to check out. The others are pretty easy to see.

To prove the last one, we should write out the enties of A and B, work out the diagonal entries of AB and the sum of them. Then work out the sum of the diagonal entries of BA and their sum. Rearraning we should see we get the same answer.

In the 2×2 we would take

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$$

(without saying what the entries are soecifically) and look at

$$AB = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & * \\ * & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix}, \quad \text{trace}(AB) = a_{11}b_{11} + a_{12}b_{21} + a_{21}b_{12} + a_{22}b_{22}$$

(where the asterisks means that we don't have to figure out what goes in those places).

$$BA = \begin{bmatrix} b_{11}a_{11} + b_{12}a_{21} & * \\ * & b_{21}a_{12} + b_{22}a_{22} \end{bmatrix}, \quad \text{trace}(BA) = b_{11}a_{11} + b_{12}a_{21} + b_{21}a_{12} + b_{22}a_{22}$$

If you look at what we got you should be abel to see that $\text{trace}(AB) = \text{trace}(BA)$. The idea is that now we know this is always going to be true for 2×2 matrices A and B, because we worked

out that the two traces are equal without having to know the values of the entries. So it has to work no matter what the numbers are that go in A and B.

All we have done is chack $\mathrm{trace}(AB) = \mathrm{trace}(BA)$ in the 2×2 case. To check it for all the sizes is not really that much more difficult but it requires a bit of nottaion to be able to keep track of what is going on.

A Appendix

Proof. (**of Theorem 5.13.2**).

(a) ⇒ (b) Assume A is invertible and A^{-1} is its inverse.

Consider the equation $A\mathbf{x} = \mathbf{0}$ where $\mathbf{x}$ is some $n \times 1$ matrix and $\mathbf{0}$ is the $n \times 1$ zero matrix. Multiply both sides by A^{-1} on the left to get

$$\begin{aligned} A^{-1}A\mathbf{x} &= A^{-1}\mathbf{0} \\ I_n\mathbf{x} &= \mathbf{0} \\ \mathbf{x} &= \mathbf{0} \end{aligned}$$

Therefore $\mathbf{x} = \mathbf{0}$ is the only possible solution of $A\mathbf{x} = \mathbf{0}$.

(b) ⇒ (c) Assume now that $\mathbf{x} = \mathbf{0}$ is the only possible solution of $A\mathbf{x} = \mathbf{0}$.

That means that when we solve $A\mathbf{x} = \mathbf{0}$ by using Gauss-Jordan elimination on the augmented matrix

$$\left[\begin{array}{c|c} & 0 \\ & 0 \\ A & \vdots \\ & 0 \end{array}\right] = [A \mid \mathbf{0}]$$

we can't end with free variables.

It is easy to see then that we must end up with a reduced row echelon form that has as many leading ones as there are unknowns. Since we are dealing with n equations in n unknowns, that means A row reduces to I_n.

(c) ⇒ (d) Suppose now that A row reduces to I_n.

Wriet down an elementary matrix for each row operation we need to row-reduce A to I_n. Say they are E_1, E_2, …, E_k. Then we know E_1A is the same as the matrix we would have after the first row operation, E_2E_1A is what we got after the second one, etc. Recall from

5.12 that all elementary matrices have inverses. So we must have

$$\begin{aligned} E_k E_{k-1} \dots E_2 E_1 A &= I_n \\ E_k^{-1} E_k E_{k-1} \dots E_2 E_1 A &= E_k^{-1} I_n \\ I_n E_{k-1} \dots E_2 E_1 A &= E_k^{-1} \\ E_{k-1} \dots E_2 E_1 A &= E_k^{-1} \\ E_{k-1}^{-1} E_{k-1} \dots E_2 E_1 A &= E_{k-1}^{-1} E_k^{-1} \\ E_{k-2} \dots E_2 E_1 A &= E_{k-1}^{-1} E_k^{-1} \end{aligned}$$

So, when we keep going in this way, we end up with

$$A = E_1^{-1} E_2^{-1} \dots E_{k-1}^{-1} E_k^{-1}$$

So we have (d) because inverses of elementary matrices are again elementary matrices.

(d) ⇒ (a) If A is a product of elementary matrices, we can use 5.12 and Theorem 5.13.1 to show that A is invertible. (The inverse of the product is the product of the inverses in the reverse order.) So we get (a).

□