MTH 211 Notes: Matrix Factorizations

1 Introduction

When you have things that can be multiplied, it is often useful to think about ways that they can be factored into simpler parts. You've definitely seen stuff like this before!

- Any integer $n \geq 2$ can be factored into a product of prime numbers.
- Factorization of polynomials can be useful in solving polynomial equations.

We can multiply matrices, so if we are given a matrix A, we could ask whether it is possible to factor A into simpler parts.

Of course, this is question is vague — what do we mean by "simpler parts"? Well, as we'll see later, this ambiguity is good. Different types of "simpler parts" will lead to different ways of factoring matrices, useful in different situations. As we go through the semester, we'll learn about a few different ways to factor matrices.

In these notes, we'll learn about a way to factor matrices that is useful in the context that we've been studying, that of solving matrix equations $A\vec{x} = \vec{b}$. In fact, there's a sense in which we've already been doing this factorization without knowing it!

2 Elementary matrices

Let's start by introducing three types of matrices that might be considered "simple". The simplest matrix is an identity matrix, but that's too simple. The ***elementary matrices*** are, in a sense, minimally more complex than identity matrices; they are matrices that you can get by performing a single row operation on an identity matrix. Because there are three kinds of row operations, there are three kinds of elementary matrices:

1. ***Rescaling matrices***: You get a rescaling matrix by multiplying a row of an identity matrix by a nonzero constant. For example,

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{3} \end{bmatrix}$$

   As transformations, rescaling matrices have the effect of rescaling one of the axes. In a real-life problem, you could use a rescaling matrix to change units in one of the variables.

2. ***Shear matrices***: You get a shear matrix by adding/subtracting a row to/from another row in an identity matrix. For example,

$$\begin{bmatrix} 1 & 0 & 4 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0 \\ -3 & 1 \end{bmatrix}.$$

   We have seen shear transformations before (at least in the two-dimensional case).

3. ***Swap matrices***: You get a swap matrix by swapping two rows in an identity matrix. For example,

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad \qquad \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

   Swap matrices have the effect of exchanging two of the variables (for example, exchanging x and y).

Collectively, matrices that are in any of the above three categories are called ***elementary matrices***. Let's see what we can do with them!

Elementary matrices are invertible

Since the elementary matrices have very specific forms, you can check that they are invertible on a case-by-case basis:

1. The inverse of a rescaling matrix is another rescaling matrix with reciprocal rescaling factor. For example, if

$$E = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \text{ then } E^{-1} = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

   Note that you can check this by actually computing EE^{-1} and $E^{-1}E$ and observing that you get I.

2. The inverse of a shear matrix is another shear matrix with the sign changed in the off-diagonal term. For example, if

$$E = \begin{bmatrix} 1 & 0 & 4 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \text{ then } E^{-1} = \begin{bmatrix} 1 & 0 & -4 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

3. A swap matrix is its own inverse!

Elementary matrices implement row operations

This is the real reason that we care about elementary matrices! Every row operation that you might do to a matrix can be implemented by multiplying by an elementary matrix! This is best illustrated with an example. Let's say you have the matrix

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

and want to do Gauss-Jordan elimination on it. Here are steps for that:

$$\begin{bmatrix} 1 & 3 & 5 \\ 2 & 8 & 4 \end{bmatrix} \xrightarrow{R2-2R1} \begin{bmatrix} 1 & 3 & 5 \\ 0 & 2 & -6 \end{bmatrix} \xrightarrow{\frac{1}{2}R2} \begin{bmatrix} 1 & 3 & 5 \\ 0 & 1 & -3 \end{bmatrix} \xrightarrow{R1-3R2} \begin{bmatrix} 1 & 0 & 14 \\ 0 & 1 & -3 \end{bmatrix}$$

Every row operation has a corresponding elementary matrix, obtained by performing the same row operation on the identity matrix with the same number of rows as A:

- The operation $R2 - 2R1$ corresponds to the shear matrix $E_1 = \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix}$.
- The operation $\frac{1}{2}R2$ corresponds to the rescaling matrix $E_2 = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$.
- The operation $R1 - 3R2$ corresponds to the shear matrix $E_3 = \begin{bmatrix} 1 & -3 \\ 0 & 1 \end{bmatrix}$.

Now here's the cool part: observe that when you multiply by these elementary matrices, it has the same effect as the row operation. Let's go through what this means in detail. Let's start with A:

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

Now let's multiply by E_1:

$$E_1 A = \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 3 & 5 \\ 2 & 8 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 5 \\ 0 & 2 & -6 \end{bmatrix}$$

Don't just read this and nod your head! Actually do the multiplication and follow along. The key observation here is that $E_1 A$ is exactly the same thing as what you get when you perform the row operation $R2 - 2R1$ to A.

Let's keep going and multiply by E_2:

$$E_2 E_1 A = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 3 & 5 \\ 0 & 2 & -6 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 5 \\ 0 & 1 & -3 \end{bmatrix}.$$

And now multiply by E_3:

$$E_3 E_2 E_1 A = \begin{bmatrix} 1 & -3 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 3 & 5 \\ 0 & 1 & -3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 14 \\ 0 & 1 & -3 \end{bmatrix}.$$

So elementary matrices give us a way to reinterpret the row operations as matrix multiplication!

OK, but why?

This is a perfectly good question to ask! The reason is that this perspective allows us to view the ***steps*** of the Gauss-Jordan elimination process as ***actual matrices*** that you can work with. There are a lot of things you can do with them, but I'll show you one thing before getting to the main topic of matrix factorization.

Remember when we looked at a matrix equation $A\vec{x} = \vec{b}$ and said "What if we could just divide by A?" Well, that doesn't always work since A might not be invertible, but the elementary matrices provide us with a next best thing. Returning to the example we did above, we can calculate the product of the elementary matrices:

$$E_3 E_2 E_1 = \begin{bmatrix} 1 & -3 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix} = \begin{bmatrix} 4 & -\frac{3}{2} \\ -1 & \frac{1}{2} \end{bmatrix}.$$

This is the magic matrix that simplifies the system!

Example: Solve the equation $A\vec{x} = \vec{b}$, where A is as above and $\vec{b} = \begin{bmatrix} 1 \\ 4 \end{bmatrix}$. Let's start by writing out the equation in full:

$$\begin{bmatrix} 1 & 3 & 5 \\ 2 & 8 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \end{bmatrix}.$$

Let's multiply both sides by the magic matrix that we just found:

$$\begin{bmatrix} 4 & -\frac{3}{2} \\ -1 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 3 & 5 \\ 2 & 8 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 4 & -\frac{3}{2} \\ -1 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 \\ 4 \end{bmatrix}$$

When we simplify, we get

$$\begin{bmatrix} 1 & 0 & 14 \\ 0 & 1 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -2 \\ 1 \end{bmatrix},$$

where the left side is in rref form, and we can quickly obtain the solution

$$\vec{x} = \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} -14 \\ 3 \\ 1 \end{bmatrix}.$$

Take a minute to think about what would happen if I asked you to solve $A\vec{x} = \vec{b}$ with the same A but a different $\vec{b}$. What parts of the calculation would stay the same? What parts would change?

3 Matrix Factorization

In the example from the previous section, we found an equation involving matrices:

$$E_3E_2E_1A = R,$$

where

$$E_1 = \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix}, \qquad E_2 = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}, \qquad E_3 = \begin{bmatrix} 1 & -3 \\ 0 & 1 \end{bmatrix}, \qquad A = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 8 & 4 \end{bmatrix}, \qquad R = \begin{bmatrix} 1 & 0 & 14 \\ 0 & 1 & -3 \end{bmatrix}.$$

Here, R is the rref form of A, and E_1, E_2, E_3 are the elementary matrices that correspond to the row operations that transform A into R.

But since the elementary matrices are invertible, we can multiply both sides of the equation by their inverses:

$$\begin{aligned} E_2E_1A &= E_3^{-1}R, \\ E_1A &= E_2^{-1}E_3^{-1}R, \\ A &= E_1^{-1}E_2^{-1}E_3^{-1}R. \end{aligned}$$

If we combine the elementary matrices into a single matrix $E = E_1^{-1}E_2^{-1}E_3^{-1}$, then we have a factorization:

$$A = ER,$$

where E is an invertible matrix, and R is an rref matrix.

A few remarks about this:

- The most important thing for you to know is that this isn't really something new. You've basically been doing this already when doing Gauss-Jordan elimination. In this factorization, R is the rref form of A. The only extra thing is that you also compute E at the same time (see the example below).
- You don't need to do any calculations to find the inverses of the elementary matrices. This is because the inverse of an elementary matrix is just the elementary matrix of the opposite row operation. ("Opposite" means that the row operations cancel each other out. For example, the opposite to $R2 - 2R1$ is $R2 + 2R1$, since subtracting 2 times row 1 cancels out with adding 2 times row 1.)
- This factorization is ***unique***. This is important because it's possible to do different sequences of row operations to put a matrix in rref form. But even if two people do Gauss-Jordan elimination in different ways, they'll end up with exactly the same E and R matrices.
- This factorization is interesting at a conceptual level, but it doesn't really improve our efficiency in solving systems of equations. The "magic matrix" approach in the previous section is better. In the next section, we'll learn about the LU factorization, which is closely related to this one but even more efficient in certain situations.

Let's return to the example from before,

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

but let's redo it from the beginning, computing E while we do Gauss-Jordan elimination. The first step is

$$\xrightarrow{R2-2R1} \begin{bmatrix} 1 & 3 & 5 \\ 0 & 2 & -6 \end{bmatrix}.$$

The opposite row operation is $R2 + 2R1$, and the corresponding elementary matrix is

$$E_1^{-1} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix}.$$

The next step is

$$\xrightarrow{\frac{1}{2}R2} \begin{bmatrix} 1 & 3 & 5 \\ 0 & 1 & -3 \end{bmatrix}.$$

The opposite row operation is $2R2$, and the corresponding elementary matrix is

$$E_2^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}.$$

Then compute

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

The last row operation is

$$\xrightarrow{R1-3R2} \begin{bmatrix} 1 & 0 & 14 \\ 0 & 1 & -3 \end{bmatrix}.$$

The opposite row operation is $R1 + 3R2$, and the corresponding elementary matrix is

$$E_3^{-1} = \begin{bmatrix} 1 & 3 \\ 0 & 1 \end{bmatrix}.$$

Then compute

$$E_1^{-1}E_2^{-1}E_3^{-1} = \begin{bmatrix} 1 & 0 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 2 & 8 \end{bmatrix}.$$

The resulting factorization is $A = ER$, where

$$E = \begin{bmatrix} 1 & 3 \\ 2 & 8 \end{bmatrix} \text{ and } R = \begin{bmatrix} 1 & 0 & 14 \\ 0 & 1 & -3 \end{bmatrix}.$$

(You should check that it works!)

The LU factorization

As mentioned earlier, the "ER factorization" is interesting but not that useful in terms of computation. But there is another factorization, the LU factorization, that is computationally efficient and is often used in real-life applications. And you pretty much already know how to do it!

You get the LU factorization by exactly the same process as the ER factorization, but instead of going all the way to rref form, you only use addition/subtraction of rows to eliminate entries below pivots, and you stop when the matrix is in echelon form.

For example, for the matrix

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

we would just need to do the one row operation

$$\xrightarrow{R2-2R1} \begin{bmatrix} 1 & 3 & 5 \\ 0 & 2 & -6 \end{bmatrix}.$$

The opposite row operation is $R2 + 2R1$, and the corresponding elementary matrix is

$$E_1^{-1} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix}.$$

If we just stop here (where we have the matrix in echelon form), then we get the factorization $A = LU$, where

$$L = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \text{ and } U = \begin{bmatrix} 1 & 3 & 5 \\ 0 & 2 & -6 \end{bmatrix}.$$

(You should check that it works!)

In this case, the matrix L has the special property of being a square matrix that is ***lower triangular*** (i.e. all the entries above the diagonal are 0) with 1's along the diagonal, and the matrix U is in echelon form.

Example: Find the LU factorization of the matrix

$$A = \begin{bmatrix} 3 & 1 & 2 \\ -9 & 0 & -4 \\ 9 & 9 & 14 \end{bmatrix}.$$

First we use the 3 as a pivot and eliminate the entries below:

$$\xrightarrow[R3-3R1]{R2+3R1} \begin{bmatrix} 3 & 1 & 2 \\ 0 & 3 & 2 \\ 0 & 6 & 8 \end{bmatrix}.$$

The opposite row operations are $R2 - 3R1$ and $R3 + 3R1$, so the corresponding matrix is

$$E_1^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix}.$$

(Because we did two row operations at once, this is a product of two elementary matrices.) Then we use the 3 in the second row as a pivot and eliminate the 6 below it:

$$\xrightarrow{R3-2R2} \begin{bmatrix} 3 & 1 & 2 \\ 0 & 3 & 2 \\ 0 & 0 & 4 \end{bmatrix}.$$

The opposite row operation is $R3 + 2R2$, so the corresponding elementary matrix is

$$E_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 2 & 1 \end{bmatrix}.$$

Now the matrix is in echelon form, so we have

$$U = \begin{bmatrix} 3 & 1 & 2 \\ 0 & 3 & 2 \\ 0 & 0 & 4 \end{bmatrix}.$$

To get L, we calculate

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

(You should check that $LU = A$ is true!)

OK, but why?

Again, a perfectly good question to ask! The LU factorization is used in situations where you have a fixed matrix A, but you want to solve the system $A\vec{x} = \vec{b}$ for lots of different vectors $\vec{b}$. One way to do this is to do Gauss-Jordan elimination on $[A|\vec{b}]$ for every vector $\vec{b}$, but you'll notice that the calculations on the left side get repeated every time.

So we could ask whether there's a way to do some of the calculations involving A ahead of time, so we can reduce the number of steps involved each time we solve $A\vec{x} = \vec{b}$. The LU factorization does exactly this. Once you have it, you can solve $A\vec{x} = \vec{b}$ fairly quickly, using the following steps:

1. First solve the equation $L\vec{y} = \vec{b}$. Because L is invertible, there is a unique solution $\vec{y}$.
2. Then solve the equation $U\vec{x} = \vec{y}$.

This works because $A\vec{x} = LU\vec{x} = L\vec{y} = \vec{b}$. I know, right??

Another question you might ask is, "Why is this better? We wanted to solve one matrix equation and when we use the LU factorization we need to solve two matrix equations." The reason is that L and U are particularly nice, making it easier to solve those systems. Since L is lower triangular, you can directly solve for x_1, substitute to solve for x_2, and so on. And since U is in echelon form, you can similarly solve from the bottom up.

Example: Use the LU factorization

$$A = \begin{bmatrix} 3 & 1 & 2 \\ -9 & 0 & -4 \\ 9 & 9 & 14 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} 3 & 1 & 2 \\ 0 & 3 & 2 \\ 0 & 0 & 4 \end{bmatrix}$$

to solve the equation $A\vec{x} = \vec{b}$, where $\vec{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

First we solve $L\vec{y} = \vec{b}$:

$$\begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

The corresponding system of equations is

$$\begin{aligned} y_1 &= 1 \\ -3y_1 + y_2 &= 1 \\ 3y_1 + 2y_2 + y_3 &= 1 \end{aligned}$$

So the first equation is $y_1 = 1$. Subbing into the second equation, we get $-3 + y_2 = 1$, so $y_2 = 4$. Subbing into the third equation, we get $3 + 8 + y_3 = 1$, so $y_3 = -10$. So

$$\vec{y} = \begin{bmatrix} 1 \\ 4 \\ -10 \end{bmatrix}.$$

Next we solve $U\vec{x} = \vec{y}$:

$$\begin{bmatrix} 3 & 1 & 2 \\ 0 & 3 & 2 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \\ -10 \end{bmatrix}.$$

The corresponding system of equations is

$$\begin{aligned} 3x_1 + x_2 + 2x_3 &= 1 \\ 3x_2 + 2x_3 &= 4 \\ 4x_3 &= -10. \end{aligned}$$

The last equation gives us $x_3 = -5/2$. Subbing into the second equation, we get $3x_2 - 5 = 4$, so $x_2 = 3$. Subbing into the first equation, we get $3x_1 + 3 - 5 = 1$, so $x_1 = 1$. So

$$\vec{x} = \begin{bmatrix} 1 \\ 3 \\ -5/2 \end{bmatrix}.$$

(You should check that it works!)

Some final remarks:

- The LU factorization only works in cases where you can get to echelon form without row swaps. If row swaps are necessary, then there's a more fancy factorization that you would use instead. We won't worry about that situation in this class.

- In the above example, the matrix A is invertible. You might think, if A is invertible, then the best way to do the calculations ahead of time is to find A^{-1}. And sometimes it is. But there are reasons you might want to avoid using A^{-1}. One is that, when done by computer, the calculations involved in finding the inverse can introduce roundoff error that accumulates with large matrices. In this case, using LU gives a more accurate result.

- Another situation where LU is preferable to the inverse is the case of a huge but "sparse" matrix (meaning that most of the entries are 0). In "big data" situations, the matrices are almost always like this, and this is what makes them manageable (you only need to store the nonzero entries). The problem is that, even if A is sparse, A^{-1} might not be sparse; however, L and U will be.