18.024–ESG Notes 2

Pramod N. Achar

Spring 2000

Inverses and determinants of matrices are closely related and come up in many different contexts. In these notes we will develop some basics of the theory of these ideas. Our initial motivation is the solution of systems of linear equations, such as

$$\begin{aligned} 2x - 5y + 4z &= -3 \\ x - 2y + z &= 5 \\ x - 4y + 6z &= 10. \end{aligned}$$

To solve the system, we want to manipulate these equations to get new equations which only contain one variable each. What do we mean by "manipulate"? We can add (multiples of) one equation to another, or we can multiply both sides of a single equation by a constant. And, of course, we are allowed to rearrange the order of the equations if we want to. This process is encoded into a shorthand notation by the method of *Gauss-Jordan elimination*. We write down just the coefficients of the above system in an augmented matrix,

$$\left(\begin{array}{ccc|c} 2 & -5 & 4 & -3 \\ 1 & -2 & 1 & 5 \\ 1 & -4 & 6 & 10 \end{array}\right),$$

to which we apply the three "moves" of Gauss-Jordan elimination:

1. Exchange two rows.
2. Add a multiple of one row to another.
3. Multiply a row by a constant.

On the other hand, we can regard the original system as a single, relatively innocuous-looking matrix equation

$$A\mathbf{x} = \mathbf{b}, \qquad \text{where } A = \begin{pmatrix} 2 & -5 & 4 \\ 1 & -2 & 1 \\ 1 & -4 & 6 \end{pmatrix}, \mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \text{ and } \mathbf{b} = \begin{pmatrix} -3 \\ 5 \\ 10 \end{pmatrix}.$$

If $A\mathbf{x} = \mathbf{b}$ were an equation of real numbers, we could solve it just be dividing both sides by A. What is an appropriate analogue of this in the world of matrices? We would like a matrix B such that $BA = I$; then, we could solve $A\mathbf{x} = \mathbf{b}$ just by multiplying both sides of the equation on the left by B, and get $\mathbf{x} = B\mathbf{b}$. Let us start making this a bit more formal.

Definition 1. Let A be a $k \times n$ matrix. An $n \times k$ matrix B is called a *left inverse* of A if $BA = I$, and a *right inverse* of A if $AB = I$.

It turns out that inverses of matrices are, in general, very poorly behaved in comparison to

$\mathbb{R}$	general matrices	square matrices
every nonzero number has an inverse	may or may not have left or right inverses	if it has a left or right inverse, it has both
there is only one inverse	there many be many inverses on one side or the other	there is only one left inverse and one right inverse, and they are equal

Before we begin thinking too hard about inverses, let us make note of a relationship between Gauss-Jordan moves and matrix multiplication.

Definition 2. The *elementary Gauss-Jordan matrices*, or just *elementary matrices*, are matrices obtained by performing exactly one Gauss-Jordan move on the identity matrix.

There are three types of Gauss-Jordan matrices, one for each type of move. Typical examples are as follows:

$$\begin{pmatrix}0&1&0\\1&0&0\\0&0&1\end{pmatrix}, \qquad \begin{pmatrix}1&0&0\\0&1&3\\0&0&1\end{pmatrix}, \qquad \begin{pmatrix}1&0&0\\0&1&0\\0&0&5\end{pmatrix}.$$

Every Gauss-Jordan move corresponds uniquely to some elementary matrix, and vice versa.

Proposition 3. *Performing a particular Gauss-Jordan move on a matrix is equivalent to multiplying it on the left by the appropriate corresponding elementary Gauss-Jordan matrix.* □

For example, if B is a 2×2 matrix, then $\left(\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right) B$ is the matrix obtained by exchanging the rows of B. We shall not write out a formal proof of this fact; the reader is invited to verify it in detail.

Now, how do we find an inverse if there is one? Consider the matrix A from our earlier example. If A has a right inverse, we can set up the following equation for it:

$$\begin{pmatrix}2&-5&4\\1&-2&1\\1&-4&6\end{pmatrix}\begin{pmatrix}a&b&c\\d&e&f\\g&h&i\end{pmatrix} = \begin{pmatrix}1&&\\&1&\\&&1\end{pmatrix}$$

This matrix equation gives nine scalar equations in nine variables. But in fact we can split the equation up by columns; for example,

$$\begin{pmatrix}2&-5&4\\1&-2&1\\1&-4&6\end{pmatrix}\begin{pmatrix}a\\d\\g\end{pmatrix} = \begin{pmatrix}1\\0\\0\end{pmatrix}.$$

The second and third columns of the putative right inverse of A yield other equations like this. Each of these three 3×3 vector equations could be solved by Gauss-Jordan elimination. In each case, the goal is to perform moves until the left-hand side of the augmented matrix looks like I. So why not combine the three equations into one larger augmented matrix and solve them all at once?

To reiterate, the procedure for finding the right inverse of A is to set up the augmented matrix

$$\left(\begin{array}{ccc|ccc}2&-5&4&1&&\\1&-2&1&&1&\\1&-4&6&&&1\end{array}\right),$$

and then perform Gauss-Jordan moves on it until the left-hand side looks like I. Then, the right-hand side will contain the nine scalars we were solving for; in other words, the right-hand side will contain the right inverse of A.

This procedure is applicable to finding right inverses of any matrix, and we can use it to find left inverses by taking the transpose of the whole equation. But we are most interested in square matrices, for which we will see that left and right inverses are the same. Henceforth, unless otherwise noted, all matrices are assumed to be square.

Let us describe the right-inverse-finding procedure once again, this time algebraically using elementary Gauss-Jordan matrices. Suppose A has a right inverse B. To find B, we set up the augmented matrix

$$(A \mid I),$$

and then we multiply on the left by elementary Gauss-Jordan matrices $E_1, \dots, E_k$, until the left-hand side of the augmented matrix becomes I. Then, the right-hand side ought to be B:

$$E_k \cdots E_1 (A \mid I) = (I \mid B).$$

Let $F = E_k \cdots E_1$. The above equation says $F(A \mid I) = (I \mid B)$. We can split up this equation of augmented matrices into two equations of square matrices:

$$FA = I \qquad\qquad FI = B$$

The second of these equations just says $F = B$; *i.e.*, that F is the right inverse of A that we were looking for. But the first equation, $FA = I$, says that F is a left inverse of A! Although we only assumed that A had a right inverse, we got a left inverse of it for free. We have just proved the following:

Lemma 4. *If A has a right inverse B, then B is also a left inverse of A.* □

Theorem 5. *If A has either a left or right inverse, then it has both. Moreover, there is exactly one left inverse and one right inverse of A, and they are equal.*

(*N.B.*: This theorem fails in every possible way for nonsquare matrices. Generally nonsquare matrices only have inverses on one side, if they have any at all; and when there is an inverse on side, there are usually infinitely many inverses on that side.)

Proof. We have already seen that if A has a right inverse B, then B is also a left inverse for A. On the other hand, suppose that A has a left inverse C:

$$CA = I.$$

We want to show that C is also a right inverse of A. But this equation says that A is a right inverse of C, so Lemma 4 tells us that A must also be a left inverse of it; *i.e.*,

$$AC = I.$$

This equation says exactly that C is a right inverse of A.

We have seen that an inverse on either side of A is also an inverse on the other side; now we need to show that there is only one such two-sided inverse. Suppose B and C are both two-sided inverses for A. In particular, they are both right inverses:

$$AB = I = AC,$$

so, using the fact that B is also a left inverse,

$$\begin{aligned} AB - AC = A(B - C) &= \mathbf{0} \\ BA(B - C) &= B\mathbf{0} \\ I(B - C) &= \mathbf{0} \\ B &= C. \end{aligned}$$

Thus, if A has any left or right inverse, then it in fact has a unique two-sided inverse. □

Definition 6. A square matrix with a left or right inverse is called *invertible*. Its unique two-sided inverse is called simply its *inverse*; the inverse of a matrix A is denoted by A^{-1}.

Again, note that the ideas of "inverse" and "invertible," unmodified by "left" and "right," do not make sense for nonsquare matrices.

Let us go back to the proof of Lemma 4 and pull out a fact that will be useful to us in discussing determinants. There, we saw that the right inverse of A was equal to a product F of elementary Gauss-Jordan matrices. Now, every invertible matrix is a right inverse to some other matrix (*viz.*, its own inverse), so we conclude the following:

Proposition 7. *Every invertible matrix can be written as a product of elementary Gauss-Jordan matrices.* □

This proposition will turn out to be very important in our investigation of determinants, so it is natural to wonder whether anything like this is true for noninvertible matrices. What happens if we try to invert a noninvertible matrix by Gauss-Jordan elimination? We get the same result that occurs whenever one tries to solve a system of linear equations that has no solution: we cannot actually obtain I on the left-hand side of the augmented matrix; the best we can do is a matrix of the form

$$K = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}. \tag{1}$$

So if A is a noninvertible square matrix, and we try to find its right inverse by performing the Gauss-Jordan moves given by elementary matrices $E_1, \ldots, E_k$, the resulting augmented matrix is described by

$$E_k \cdots E_1(A \mid I) = (K \mid ?).$$

(The right-hand side of the resulting augmented matrix is not really meaningful to us, since A does not actually have a right inverse.) The left-hand parts of the augmented matrices give

$$E_k \cdots E_1 A = K,$$

or, using the following lemma,

$$A = E_1^{-1} \cdots E_k^{-1} K. \tag{2}$$

Lemma 8. *Every elementary Gauss-Jordan matrix is invertible. Moreover, the inverse of an elementary Gauss-Jordan matrix is again an elementary Gauss-Jordan matrix; the latter corresponds to undoing to Gauss-Jordan move associated with the former.*

Proof. Left as an exercise for the reader. □

Equation (2) gives us the following analogue of Proposition 7:

Proposition 9. *Every noninvertible matrix can be written as a product of elementary matrices, followed by a matrix of the form shown in Equation (1).* □

Now, we are finally ready to start talking about determinants. We take an axiomatic approach to determinants, as we did with the trigonometric functions in the first semester. We suppose that there is a function

$$\det : \{n \times n \text{ matrices}\} \to \mathbb{R}$$

having the following four properties (here we regard the rows of a matrix as a list of n vectors in $\mathbb{R}^n$):

1. $\det \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ c\mathbf{a}_i \\ \vdots \\ \mathbf{a}_n \end{pmatrix} = c \det \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_i \\ \vdots \\ \mathbf{a}_n \end{pmatrix}$.

2. If $\det \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_i + \mathbf{b}_i \\ \vdots \\ \mathbf{a}_n \end{pmatrix} = \det \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_i \\ \vdots \\ \mathbf{a}_n \end{pmatrix} + \det \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{b}_i \\ \vdots \\ \mathbf{a}_n \end{pmatrix}$.

3. If two rows $\mathbf{a}_i$ and $\mathbf{a}_j$ of a matrix are equal, then its determinant is zero.

4. $\det I = 1$.

It can immediately be seen how the determinant changes in response to Gauss-Jordan moves on a matrix. The following proposition was discussed in detail in class, and its proof will not be repeated here.

Proposition 10. *(a) If two rows are exchanged, the determinant is multiplied by -1.*

(b) If a multiple of one row is added to another, the determinant is unchanged.

(c) If a row is multiplied by a constant c, the determinant is also multiplied by c. □

In particular, since the elementary Gauss-Jordan matrices are obtained by performing Gauss-Jordan moves on I, whose determinant is 1, we can find the determinants of the elementary Gauss-Jordan matrices. Since performing a Gauss-Jordan move is the same as multiplying on the left by an elementary matrix, and since the change in determinant after such a move is precisely the determinant of the corresponding elementary matrix, we have the following:

Lemma 11. *If E is an elementary Gauss-Jordan matrix, and B is any matrix, then $\det EB = \det E \det B$.* □

Now, Proposition 7 tells us that any invertible matrix is a product of elementary ones. Suppose $A = E_1 \cdots E_k$ is invertible, where the E_i's are elementary matrices. Then

$$\begin{aligned}
\det AB &= \det E_1 \cdots E_k B \\
&= \det E_1 \det E_2 \cdots E_k B \\
&= \det E_1 \cdots \det E_k \det B \\
&= \det E_1 \cdots \det E_{k-1} E_k \det B \\
&= \det E_1 \cdots E_k \det B \\
&= \det A \det B.
\end{aligned}$$

(Here we used the preceding lemma several times to pull elementary matrices in or out of a determinant.) We have just shown:

Lemma 12. *If A is invertible and B is any matrix, then $\det AB = \det A \det B$.* □

What if A is noninvertible? The simplest kind of noninvertible matrix is that of the type shown in Equation (1). The matrix K shown there has $\mathbf{0}$ as its final row, so of course if we multiply that row by 0, the matrix is unchanged. On the other hand, multiplying any row by 0 should multiply its determinant by 0 as well. This is only possible if $\det K = 0$ to begin with.

Theorem 13. *A matrix is invertible if and only if it has nonzero determinant.*

Proof. If A is invertible, it is a product of elementary matrices, and its determinant is the product of the determinants of those elementary matrices, each of which is nonzero, so $\det A \neq 0$.

On the other hand, if A is noninvertible, it is a product of elementary matrices with a matrix of the form of K, as discussed above. Again $\det A$ is the product of the determinants of these matrices, but since $\det K = 0$, we have $\det A = 0$ as well. □

Proposition 14. *If A is not invertible, neither is AB for any matrix B.*

Proof. If AB had an inverse C, then $(AB)C = I$. But this implies $A(BC) = I$, which says that BC is an inverse of A. Since we have assumed A to be noninvertible, this is a contradiction. □

Theorem 15. $\det AB = \det A \det B$.

Proof. We have already proved this when A is invertible, so we only need to prove it when A is not invertible. In this case, AB is also not invertiable, so we have $\det AB = 0 = \det A$. It follows that $\det AB = \det A \det B$. □

Without noticing it, we have already proved that the determinant function must be unique. We have seen that the axioms imply the values of the determinants of the elementary matrices and the noninvertible matrix K; we know that every matrix can be written as a product of these types of matrices; and we know how to treat determinants of products. In other words, from Propositions 7 and 9 and Theorem 15, it follows that:

Proposition 16. *There is at most one function satisfying the axioms for determinant. In other words, provided that the determinant function exists, it is unique.* □

How might it fail to exist? It is not obvious *a priori* how to write down a function satisfying the four axioms we have given, but nevertheless we almost have an algorithm for computing the determinant: just do Gauss-Jordan elimination to find the kind of decomposition described in Proposition 7 or 9, and then multiply together the determinants of the elementary matrices thus obtained. How could a function not exist when we have a way to compute it?

The reason is that the Gauss-Jordan elimination process does not yield a unique answer. When we perform Gauss-Jordan elimination, we have a great deal of choice in what moves to perform in what order. If we have two *different* ways of writing A as a product of elementary matrices, which one do we use to compute the determinant? Or do they give the same answer? In other words, the question of whether the determinant function exists is a question of whether different sequences of Gauss-Jordan moves give the same answer.

We will answer the question by giving an explicit formula for computing determinants and showing that it satisfies the four axioms. It follows from what we have proved so far that *any* way of writing a matrix as a product of elementary ones must give the same answer as that given by the formula. In practice, however, especially for very large matrices, this formula is unwieldy. Engineers and computers in real life never use this formula; they use Gauss-Jordan elimination.[1]

The formula in question is called "expansion by minors." A minor is a smaller matrix obtained from a given one by deleting some rows and columns. In particular, if A is an $n \times n$ matrix, let A_{ij} be the $(n-1) \times (n-1)$ minor obtained by deleting the ith row and jth column. Furthermore, let a_{ij} be the entry in row i, column j of A. The formula in the following theorem is called "expansion by minors along the first column":

Theorem 17. *The determinant of a* 1×1 *matrix is given by* $\det(a) = a$*, and the determinants of larger matrices are given by*

$$\begin{aligned} \det A &= \sum_{i=1}^{n} (-1)^{i+1} a_{i1} \det A_{i1} \\ &= a_{11} \det A_{11} - a_{12} \det A_{12} + \cdots + (-1)^{n+1} a_{n1} \det A_{n1}. \end{aligned} \tag{3}$$

Proof. We proceed by induction on the size of the matrix. For 1×1 matrices, it is utterly obvious that the formula $\det(a) = a$ satisfies axioms 1, 2, and 4; and axiom 3 is inapplicable. Now we assume that there exists a determinant function for $(n-1) \times (n-1)$ matrices, satisfying the four axioms. That is, we are allowed to use the axioms when dealing with determinants of minors, but we are trying to prove that they hold for the determinant of the whole matrix.

Suppose first that we multiply row i of A by a constant c. What happens in Equation (3)? In each term $a_{k1} \det A_{k1}$ with $k \neq i$, the minor A_{k1} contains part of row i, so we know (since it is an $(n-1) \times (n-1)$ matrix) that its determinant gets multiplied by c. What about the term $a_{i1} \det A_{i1}$? Here the minor does not contain row i, but the coefficient itself has been multiplied by c. Thus, every term of (3) gets multiplied by c, so this formula satisfies Axiom 1.

Now, suppose row i of A contains a sum of two vectors,

$$\mathbf{a}_i + \mathbf{b}_i = (a_{i1}, a_{i2}, \ldots, a_{in}) + (b_{i1}, b_{i2}, \ldots, b_{in}).$$

We want to split up formula (3) into two determinants. Again, in every term $a_{k1} \det A_{k1}$ with $k \neq 1$, the minor A_{k1} contains part of the ith row of A, and its determinant can be split up into two determinants using

[1] If I am remembering correctly, computing the determinant of an $n \times n$ matrix by Gauss-Jordan elimination takes $O(n^2)$ time, whereas computing it by the formula of expansion by minors takes $O(n!)$ time.

Axiom 2. On the other hand, the ith term is $(a_{i1} + b_{i1}) \det A_{i1}$: here the minor does not contain row i of A, but the coefficient can be split up to yield two terms $a_{i1} \det A_{i1} + b_{i1} \det A_{i1}$. Every term in (3) can be split into two terms, therefore; and these can be regrouped to give the sum of two determinant expressions as demanded by Axiom 2. Thus, Axiom 2 is satisfied.

Next, suppose that row i and row j of A are equal. Every minor A_{k1}, except $k = i$ and $k = j$, will also have a repeated row, and hence determinant 0. So nearly all the terms in (3) vanish; only the ith and jth remain:

$$detA = (-1)^{i+1} a_{i1} \det A_{i1} + (-1)^{j+1} a_{j1} \det A_{j1}. \tag{4}$$

Let us assume, without loss of generality, that $i < j$. How are the minors A_{i1} and A_j1 related? They are obtained by deleting rows i and j, respectively, as well as column 1, from A. Since rows i and j are the same, A_{i1} and A_{j1} contain the same entries, although the rows are rearranged somewhat. To get A_{i1} from A_{j1}, we would remove the ith row of A_{j1}, scoot rows $i+1, \ldots, j-1$ up one row, and then insert the removed ith row into the now-empty $(j-1)$th row. For example, consider the 5×5 matrix

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

The minors A_{51} and A_{21} are related as follows: we remove row 2, scoot rows 3 and 4 up one position, and then insert the deleted row into row 4.

$$A_{51} = \begin{pmatrix} 1 & 8 & 3 & 6 \\ 6 & 5 & 3 & 0 \\ 5 & 8 & 0 & 7 \\ 8 & 7 & 4 & 2 \end{pmatrix} \mapsto \begin{pmatrix} 1 & 8 & 3 & 6 \\ & & & \\ 5 & 8 & 0 & 7 \\ 8 & 7 & 4 & 2 \end{pmatrix} \mapsto \begin{pmatrix} 1 & 8 & 3 & 6 \\ 5 & 8 & 0 & 7 \\ 8 & 7 & 4 & 2 \\ & & & \end{pmatrix} \mapsto \begin{pmatrix} 1 & 8 & 3 & 6 \\ 5 & 8 & 0 & 7 \\ 8 & 7 & 4 & 2 \\ 6 & 5 & 3 & 0 \end{pmatrix} = A_{21}.$$

Another way of looking at this is that we start A_{j1}, then exchange rows i and $i+1$, then exchange rows $i+1$ and $i+2$, *etc.*, until at last we exchange rows $j-2$ and $j-1$. This has the effect of gradually pushing what was originally row i down into position $j-1$ while simultaneously scooting up the intermediate rows by one position each. In all, we have to perform $j-i-1$ row exchanges to change A_{j1} into A_{i1}. These are $(n-1) \times (n-1)$ matrices, so we know that exchanging rows multiplies the determinant by -1. Therefore,

$$\det A_{i1} = (-1)^{j-i-1} \det A_{j1}.$$

Plugging this back into (4), we get

$$\begin{aligned} \det A &= (-1)^{i+1} a_{i1} (-1)^{j-i-1} \det A_{j1} + (-1)^{j+1} a_{j1} \det A_{j1} \\ &= ((-1)^j + (-1)^{j+1}) a_{j1} \det A_{j1} \\ &= 0. \end{aligned}$$

(Recall that $a_{i1} = a_{j1}$.) This is exactly what we wanted to show; we have established that Axiom 3 is satisfied.

Finally, let us compute $\det I$. All the a_{i1} are 0 except a_11, so the formula in (3) has just one nonzero term: $\det I = a_{11} \det A_{11}$. Moreover, $a_{11} = 1$, and the minor A_{11} is just a smaller identity matrix, so we know that it has determinant 1. Hence $\det I = 1$, and Axiom 4 is satisfied as well.

We conclude that the determinant function exists and is given by the formula in (3). □

Whew!