Lecture 3: LINEAR ALGEBRA FOUNDATIONS *

10-606 MATHEMATICAL FOUNDATIONS FOR MACHINE LEARNING

1 Linear Algebra

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

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

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

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

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

1.1 Basic Notation

Throughout this course, we will use the following notation:

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

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

*Compiled on Friday 23rd June, 2023 at 15:43

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

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

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

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

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

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

  - Note that these definitions are ambiguous, e.g., the a_1 and a_1^T in the previous two definitions are *not* the same vector. Typically the meaning of the notation will be obvious from its use.

2 Matrix Multiplication

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

$$C = AB \in \mathbb{R}^{m \times p} \text{ where } C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}.$$

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

2.1 Vector-Vector Products

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

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

Note that it is always the case that $x^T y = y^T x$.

Given vectors $x \in \mathbb{R}^m$, $y \in \mathbb{R}^n$ (they no longer have to be the same size), xy^T is called the *outer product* of the vectors. It is a matrix whose entries are given by $(xy^T)_{ij} = x_i y_j$ i.e.,

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

2.2 Matrix-Vector Products

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

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

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

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

Alternatively, if we write A in column form, we see that,

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

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

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

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

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

Lastly, expressing A in terms of rows we get another representation of the vector-matrix product,

$$\begin{aligned} y^T &= \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} \text{---} & a_1^T & \text{---} \\ \text{---} & a_2^T & \text{---} \\ & \vdots & \\ \text{---} & a_m^T & \text{---} \end{bmatrix} \\ &= x_1 \begin{bmatrix} \text{---} & a_1^T & \text{---} \end{bmatrix} + x_2 \begin{bmatrix} \text{---} & a_2^T & \text{---} \end{bmatrix} + ... + x_n \begin{bmatrix} \text{---} & a_n^T & \text{---} \end{bmatrix}. \end{aligned}$$

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

2.3 Matrix-Matrix Products

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

First we can view matrix-matrix multiplication as a set of vector-vector products. The most obvious viewpoint, which follows immediately from the definition, is that the i, j entry of C is equal to the inner product of the ith row of A and the jth row of B. Symbolically, this looks like

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

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

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

Put another way, AB is equal to the sum, over all i, of the outer product of the ith column of A and the ith row of B. Since, in this case, $a_i \in \mathbb{R}^m$ and $b_i \in \mathbb{R}^p$, the dimension of the outer product $a_i b_i^T$ is $m \times p$, which coincides with the dimension of C.

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

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

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

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

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

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

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

- Matrix multiplication is associative: $(AB)C = A(BC)$.
- Matrix multiplication is distributive: $A(B + C) = AB + AC$.
- Matrix multiplication is, in general, *not* commutative; that is, it can be the case that $AB \neq BA$.

3 Operations and Properties

In this section we present several operations and properties of matrices and vectors.

3.1 The Transpose

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

$$A^T \in \mathbb{R}^{n\times m}, (A^T)_{ij} = A_{ji} \ .$$

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

The following properties of transposes are easily verified:

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

Knowledge check

1. **Select all that apply**: Suppose $A \in \mathbb{R}^{2\times 3}$ and $B \in \mathbb{R}^{2\times 5}$. Which of the following are valid matrix multiplications?
   a) AB
   b) BA
   c) $A^T B$
   d) $B^T A$
   e) None of the above
   - **Answer**: c and d; in order for a matrix multiplication to be valid, the number of columns in the first matrix must equal the number of rows in the second matrix.
2. **Numerical answer**: For each of the options you selected above, what is the size of the resulting matrix multiplication i.e., how many rows and columns does the product have?
   - **Answer**: $A^T B \in \mathbb{R}^{3\times 5}$ and $B^T A \in \mathbb{R}^{5\times 3}$

3.2 Symmetric Matrices

A square matrix $A \in \mathbb{R}^{n\times n}$ is *symmetric* if $A = A^T$. It is *anti-symmetric* if $A = -A^T$. It is easy to show that for any matrix $A \in \mathbb{R}^{n\times n}$, the matrix $A + A^T$ is symmetric and the matrix $A - A^T$ is anti-symmetric. From this it follows that any square matrix $A \in \mathbb{R}^{n\times n}$ can be represented as a sum of a symmetric matrix and an anti-symmetric matrix, since

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

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

3.3 The Trace

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

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

The trace has the following properties:

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

3.4 Norms

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

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

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

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

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

Other examples of norms are the ℓ_1 norm,

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

and the ℓ_∞ norm,

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

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

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

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

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

3.5 The Determinant

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

Given a matrix

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

consider the set of points $S \subseteq \mathbb{R}^n$ formed by taking all possible linear combinations of the row vectors $a_1, \ldots, a_n \in \mathbb{R}^n$ of A, where the coefficients of the linear combination are all between 0 and 1:

$$S = \{v \in \mathbb{R}^n : v = \sum_{i=1}^{n} c_i a_i, 0 \leq c_i \leq 1 \forall i \in \{1, \ldots, n\}\}$$

(this is a bit of a strange definition but will make more sense once we define the notion of the *span* of a matrix, which we will do so later). The absolute value of the determinant of A, it turns out, is a measure of the "volume" of the set S (admittedly, we have not actually defined what we mean by "volume" here, but hopefully the intuition should be clear enough: when $n = 2$, our notion of volume corresponds to the area of S in the Cartesian plane and when $n = 3$, volume corresponds with our usual notion of the size a three-dimensional object).

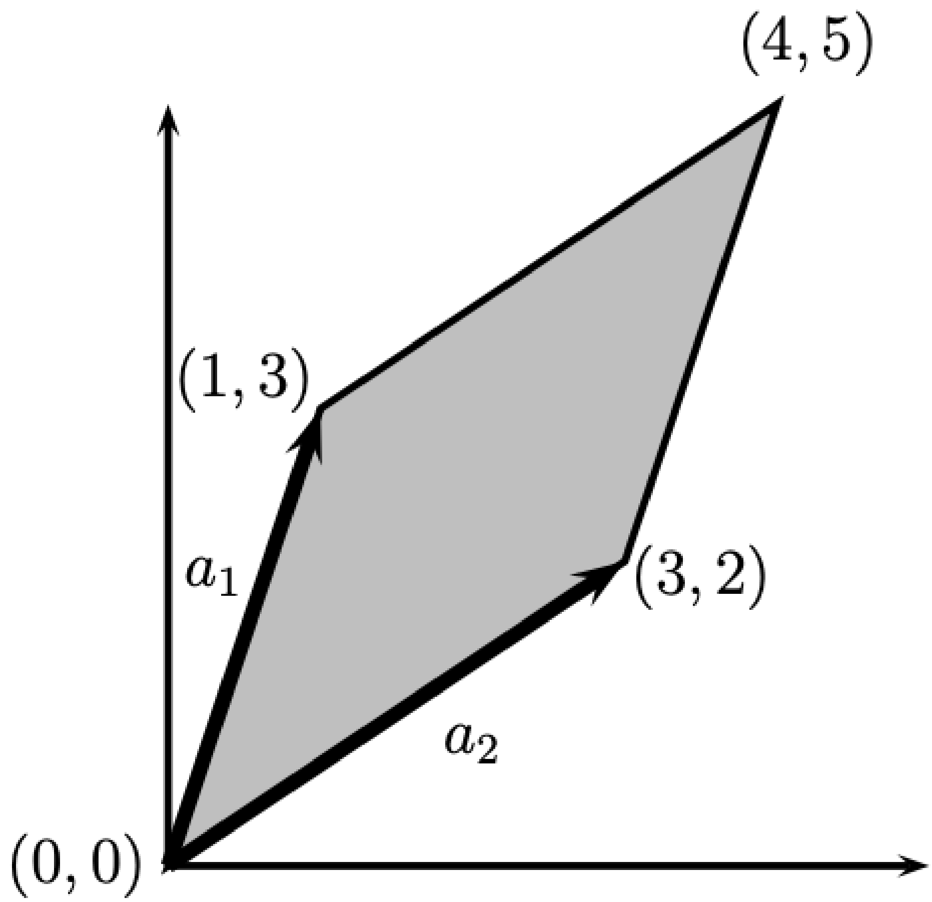
For example, consider the 2×2 matrix,

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

Here the rows of this matrix are

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

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



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

Algebraically, the determinant satisfies the following properties (from which all other properties follow, including the general formula which we shall provide shortly):

1. The determinant of the *identity* matrix I_n, an $n \times n$ matrix where all of the diagonal elements are 1 and all other elements are 0, is 1 i.e., $|I| = 1$. Geometrically, I_n corresponds to the n-dimensional hypercube which has volume 1.

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

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

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

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

These properties, however, give very little intuition about the nature of the determinant, so we now list several properties that follow from the three properties above:

- For $A \in \mathbb{R}^{n\times n}$, $|A| = |A^T|$.
- For $A, B \in \mathbb{R}^{n\times n}$, $|AB| = |A||B|$.

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

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

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

$$\begin{aligned} |[a_{11}]| &= a_{11} \\ \left|\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\right| &= a_{11}a_{22} - a_{12}a_{21} \end{aligned}$$

Knowledge check

1. **Math**: Use the recursive formula and the formula for the determinant of a 2×2 matrix, both provided above, to derive the formula for the determinant of a 3×3 matrix:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

   - **Answer**: Using the second recursive definition for $|A|$ given above (the one that sums over j) and arbitrarily picking $i = 1$, we can explicitly write out that sum as

$$\begin{aligned} &\sum_{j=1}^{n} (-1)^{i+j} a_{ij} |A_{\setminus i,\setminus j}| \\ =&(-1)^{1+1} a_{11} |A_{\setminus 1,\setminus 1}| + (-1)^{1+2} a_{12} |A_{\setminus 1,\setminus 2}| + (-1)^{1+3} a_{13} |A_{\setminus 1,\setminus 3}| \\ =&a_{11} \left| \begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix} \right| - a_{12} \left| \begin{bmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{bmatrix} \right| + a_{13} \left| \begin{bmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \right| \end{aligned}$$

     Plugging in the formula for the 2×2 determinants gives

$$a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31}$$