

UNIVERSAL IDENTITIES

KEITH CONRAD

1. INTRODUCTION

We want to describe an idea that reduces the verification of algebraic identities valid in all commutative rings to the verification in the complex numbers, where special tools (from linear algebra, geometry, or analysis) are available.

What is meant by an algebraic identity valid in all commutative rings? To take a simple example, consider the multiplicativity of sums of two squares:

$$(1.1) \quad (a^2 + b^2)(c^2 + d^2) = (ac - bd)^2 + (ad + bc)^2.$$

A direct calculation shows that (1.1) holds in every commutative ring. That is, (1.1) is true when the 4 parameters $a, b, c,$ and d are arbitrary elements of an arbitrary commutative ring. But there is another way of thinking about (1.1), as a polynomial identity in 4 indeterminates A, B, C, D with coefficients in \mathbf{Z} . That is, (1.1) says

$$(1.2) \quad (A^2 + B^2)(C^2 + D^2) = (AC - BD)^2 + (AD + BC)^2$$

in $\mathbf{Z}[A, B, C, D]$ where the parameters are indeterminates. Actually, (1.2) is just a particular case of (1.1) using $a = A, b = B, c = C,$ and $d = D$ in the polynomial ring $\mathbf{Z}[A, B, C, D]$.

Another instance of an algebraic identity valid in all commutative rings is multiplicativity of determinants. In the 2×2 case this says

$$\det \left(\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix} \right) = \det \begin{pmatrix} a & b \\ c & d \end{pmatrix} \det \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix},$$

or equivalently

$$(aa' + bc')(cb' + dd') - (ab' + bd')(ca' + dc') = (ad - bc)(a'd' - b'c').$$

A particular case of this is $a = A, b = B, \dots, d' = D'$ in $\mathbf{Z}[A, B, C, D, A', B', C', D']$.

In Section 2 we will describe how algebraic identities that make sense in all commutative rings can be proved by working only in \mathbf{C} . This is a really significant idea! In Section 3 we'll state three identities about determinants that will be proved by reduction to the complex case, including the Cayley-Hamilton theorem. Proofs will be given in Section 4, while Section 5 discusses some interesting consequences of the Cayley-Hamilton theorem.

We will be dealing with multivariable polynomials, and will use an abbreviated notation for them. Rather than writing

$$f(X_1, \dots, X_n) = \sum_{i_1=0}^{d_1} \sum_{i_2=0}^{d_2} \cdots \sum_{i_n=0}^{d_n} c_{i_1, \dots, i_n} X_1^{i_1} \cdots X_n^{i_n}$$

we write

$$f(X_1, \dots, X_n) = \sum_{i_1, \dots, i_n} c_{i_1, \dots, i_n} X_1^{i_1} \cdots X_n^{i_n}$$

since the degree of f in each of X_1, \dots, X_n often won't matter.

It is important that we can think about a polynomial in several indeterminates as a polynomial in one indeterminate whose coefficients are polynomials in the other indeterminates. For example, $f(X, Y) = 3X^4Y + X^2Y^3 + XY^3 + X - 7$ is cubic in Y and quartic in X :

$$f(X, Y) = (X^2 + X)Y^3 + (3X^4)Y + (X - 7) = (3Y)X^4 + (Y^3)X^2 + (Y^3 + 1)X - 7.$$

2. REDUCTION TO THE COMPLEX CASE

An identity in all commutative rings such as (1.1) includes as a special case an identity (1.2) of polynomials with coefficients in \mathbf{Z} , but this special case in turn implies the general case, since we can substitute elements of a commutative ring for the indeterminates.

Theorem 2.1. *For a commutative ring R and $a_1, \dots, a_n \in R$, the substitution map sending each*

$$f(X_1, \dots, X_n) = \sum_{i_1, \dots, i_n} c_{i_1, \dots, i_n} X_1^{i_1} \cdots X_n^{i_n}$$

in $\mathbf{Z}[X_1, \dots, X_n]$ to its value

$$f(a_1, \dots, a_n) = \sum_{i_1, \dots, i_n} c_{i_1, \dots, i_n} a_1^{i_1} \cdots a_n^{i_n}$$

at a_1, \dots, a_n is a ring homomorphism $\mathbf{Z}[X_1, \dots, X_n] \rightarrow R$.

Here f varies while the values a_1, \dots, a_n are fixed.

Proof. There is only one ring homomorphism $\mathbf{Z} \rightarrow R$, so the integral coefficients of the polynomial have only one possible meaning in R . The definitions of addition and multiplication in $\mathbf{Z}[X_1, \dots, X_n]$ show that replacing the indeterminates in polynomials by particular values in a commutative ring R is a ring homomorphism $\mathbf{Z}[X_1, \dots, X_n] \rightarrow R$. Details are left to the reader. \square

A slogan for this idea is “substitution (or specialization) is a ring homomorphism.”

Example 2.2. Substituting four arbitrary elements of a commutative ring R into both sides of the polynomial identity (1.2) recovers (1.1).

Example 2.3. Here is an example where substitution does not behave well. In $\mathbf{Z}[X, Y]$, $X^2 - Y^2 = (X + Y)(X - Y)$. If we replace X with $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and Y with $\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$ then $X^2 - Y^2$ goes to $\begin{pmatrix} 0 & 0 \\ -2 & 0 \end{pmatrix}$ while $(X + Y)(X - Y)$ goes to $\begin{pmatrix} -1 & 2 \\ -2 & 1 \end{pmatrix}$, which is different. The reason this does not violate Theorem 2.1 is that the substitution we made involves noncommuting matrices. In fact, for a polynomial $f(X, Y)$, the meaning of $f(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix})$ is ambiguous since the way we write $f(X, Y)$ as a polynomial (e.g., should a term X^2Y be YX^2 or XYX ?) affects its value when we replace X and Y with $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$. Substituting noncommuting elements into a polynomial is well-defined only if we make an agreement in advance about how the polynomial is written. Substitution of commuting elements into a polynomial doesn't have such problems.

Remark 2.4. There are no problems as in Example 2.3 if we substitute matrices into a polynomial when the matrices commute. That is, Theorem 2.1 is okay when R is a noncommutative ring as long as $a_1, \dots, a_n \in R$ commute. In particular, a polynomial identity in one variable remains an identity if we replace the variable with a matrix (and make the constant term a constant multiple of the identity matrix). For instance, $X^2 - 1 = (X + 1)(X - 1)$ in $\mathbf{Z}[X]$ and $M^2 - I_n = (M + I_n)(M - I_n)$ for every $n \times n$ matrix M over a commutative ring.

Suppose we have a polynomial identity in $\mathbf{Z}[X_1, \dots, X_n]$, say

$$f(X_1, \dots, X_n) = g(X_1, \dots, X_n),$$

which by definition means coefficients of the same monomials on both sides are the same integer. An example is (1.2). By Theorem 2.1, substituting values from a commutative ring for the X_i 's maintains the equality because substitution is a ring homomorphism. Thus $f(a_1, \dots, a_n) = g(a_1, \dots, a_n)$, where the a_i 's are arbitrary elements of a commutative ring: a polynomial identity with integral coefficients remains true under specialization into every commutative ring. This is the key idea we will use repeatedly.

Remark 2.5. If we want to prove there is no polynomial identity of a certain kind then we may be able to use values in \mathbf{Z} to find a counterexample. For instance, there is no analogue of (1.2) for sums of three squares. Indeed, assume there is a polynomial identity

$$(2.1) \quad (A^2 + B^2 + C^2)(A'^2 + B'^2 + C'^2) = f^2 + g^2 + h^2$$

for indeterminates $A, B, C, A', B',$ and C' , and $f, g,$ and h in $\mathbf{Z}[A, B, C, A', B', C']$. From (2.1) we get a similar formula for products of sums of three squares in every commutative ring by specializing the 6 indeterminates A, B, \dots, C' to 6 arbitrary elements of a commutative ring. So (2.1) implies that sums of three squares are closed under multiplication in every commutative ring, but this false in the ring \mathbf{Z} : 3 and 5 are sums of three squares in \mathbf{Z} (we include 0 as a square), but their product 15 is not. Therefore a polynomial identity of the form (2.1) does not exist!

To prove an algebraic identity holds in all commutative rings, why is it useful to view it as a polynomial identity with coefficients in \mathbf{Z} ? After all, the actual algebraic calculations that are used to verify (1.2) are the same as the ones used to verify (1.1), so there is no advantage gained by viewing (1.1) as the polynomial identity (1.2). But there are much more complicated identities, such as in linear algebra, where the polynomial identity viewpoint is really useful thanks to the following theorem involving complex numbers. Notice the topological hypothesis in the theorem.

Theorem 2.6. *Let $f(X_1, \dots, X_n)$ and $g(X_1, \dots, X_n)$ be in $\mathbf{C}[X_1, \dots, X_n]$. If f and g are equal functions on a nonempty open set in \mathbf{C}^n then $f = g$ in $\mathbf{C}[X_1, \dots, X_n]$.*

Proof. We reformulate the theorem in terms of $f - g$: if a polynomial in $\mathbf{C}[X_1, \dots, X_n]$ vanishes on an open set in \mathbf{C}^n then the polynomial is 0 in $\mathbf{C}[X_1, \dots, X_n]$ (that is, all of its coefficients are 0). If $n = 1$ then the proof is easy: a polynomial in $\mathbf{C}[X]$ that vanishes on a nonempty open set in \mathbf{C} has an infinite number of roots. Since polynomials in $\mathbf{C}[X]$ other than 0 have finitely many roots, only the zero polynomial vanishes on a nonempty open set in \mathbf{C} . Now assume $n \geq 1$ and the only polynomial in $\mathbf{C}[X_1, \dots, X_n]$ vanishing on a nonempty open set in \mathbf{C}^n is the zero polynomial. For a polynomial $f(X_1, \dots, X_{n+1})$ vanishing on a nonempty open set in \mathbf{C}^{n+1} , we will prove $f = 0$ as a polynomial (*i.e.*, all its coefficients are 0) by reduction to the previous case of polynomials in n variables. Write f as a polynomial in X_{n+1} with coefficients that are polynomials in X_1, \dots, X_n :

$$f(X_1, \dots, X_{n+1}) = \sum_{i=0}^d c_i(X_1, \dots, X_n) X_{n+1}^i,$$

where $c_i \in \mathbf{C}[X_1, \dots, X_n]$. We will show each c_i equals 0 in $\mathbf{C}[X_1, \dots, X_n]$, so $f = 0$.

Let $U \subset \mathbf{C}^{n+1}$ be a nonempty open set in \mathbf{C}^{n+1} where f vanishes: if $(z_1, \dots, z_{n+1}) \in U$ then $f(z_1, \dots, z_{n+1}) = 0$. From the topology of \mathbf{C}^{n+1} , U contains (around each point inside

it) a direct product $U_1 \times \cdots \times U_{n+1}$ where each U_i is a nonempty open set in \mathbf{C} . Pick a point $(z_1, \dots, z_{n+1}) \in U_1 \times \cdots \times U_{n+1}$ and consider the following one-variable polynomial derived from specializing all but the last variable in $f(X_1, \dots, X_{n+1})$:

$$g(X) = f(z_1, \dots, z_n, X) = \sum_{i=0}^d c_i(z_1, \dots, z_n) X^i \in \mathbf{C}[X].$$

For all $z \in U_{n+1}$, $(z_1, \dots, z_n, z) \in U$, so $g(z) = 0$. Therefore $g(X)$ vanishes on a nonempty open set U_{n+1} in \mathbf{C} , so all coefficients of $g(X)$ equal 0 by the base case $n = 1$. Thus

$$(2.2) \quad c_i(z_1, \dots, z_n) = 0$$

for $i = 0, \dots, d$ and $(z_1, \dots, z_n) \in U_1 \times \cdots \times U_n$. Since $U_1 \times \cdots \times U_n$ is a nonempty open set in \mathbf{C}^n , by induction on n each c_i is the zero polynomial in $\mathbf{C}[X_1, \dots, X_n]$, so $f = 0$ in $\mathbf{C}[X_1, \dots, X_{n+1}]$. This concludes the proof. \square

Theorem 2.6 is true with \mathbf{C} replaced by \mathbf{R} , using exactly the same proof (open sets in \mathbf{R}^n replace open sets in \mathbf{C}^n), but the applications we have in mind will require us to work in \mathbf{C} , so we only stated the theorem that way. However, working in \mathbf{R} gives us a picture of why the theorem is true geometrically. If $f(X, Y) \in \mathbf{R}[X, Y]$ is a nonzero polynomial then the equation $f(x, y) = 0$ usually traces out a curve in the plane, which is locally one-dimensional and certainly contains no open set of \mathbf{R}^2 . (A curve in the plane contains no open ball.) So if $f(X, Y) \in \mathbf{R}[X, Y]$ and $f(x, y) = 0$ for all (x, y) in some nonempty open set in \mathbf{R}^2 , then $f(X, Y) = 0$. This reasoning extends to more than 2 variables and to complex coefficients if the reader has experience with differential or algebraic geometry.

Combining Theorems 2.1 and 2.6, we have the following procedure for reducing the verification of an algebraic identity in all commutative rings (such as (1.2)) to its verification as a polynomial identity over \mathbf{C} . We call it the “method of universal identities”.

- (1) Express the identity as $f(a_1, \dots, a_n) = g(a_1, \dots, a_n)$ for f and g in $\mathbf{Z}[X_1, \dots, X_n]$ and a_1, \dots, a_n running over all elements of an arbitrary commutative ring.
- (2) Verify $f(z_1, \dots, z_n) = g(z_1, \dots, z_n)$ as (z_1, \dots, z_n) runs over an open set in \mathbf{C}^n .
- (3) By Theorem 2.6, $f(X_1, \dots, X_n) = g(X_1, \dots, X_n)$ in $\mathbf{Z}[X_1, \dots, X_n]$.
- (4) By Theorem 2.1, $f(a_1, \dots, a_n) = g(a_1, \dots, a_n)$, where the a_i 's are taken from a commutative ring. That is, our identity is true in all commutative rings.

The identities we will prove below by this method involve determinants of matrices. We will reduce the proof of such identities to the proof for matrices in $M_n(\mathbf{C})$, which can be thought of as \mathbf{C}^{n^2} in a natural way. The topology of $M_n(\mathbf{C})$ arising from its identification with \mathbf{C}^{n^2} is the one where matrices are considered close when they are entrywise close. For example, a neighborhood of $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is all matrices $\begin{pmatrix} a+\varepsilon_1 & b+\varepsilon_2 \\ c+\varepsilon_3 & d+\varepsilon_4 \end{pmatrix}$ with small ε_i .

To apply Theorem 2.6 to polynomials in n^2 variables, thought of as functions of the entries of $n \times n$ matrices, we need to use open sets in $M_n(\mathbf{C})$. The specific examples of open sets we will use are described in the next two theorems.

Theorem 2.7. *The group $\mathrm{GL}_n(\mathbf{C})$ is open in $M_n(\mathbf{C})$.*

Proof. The determinant function $\det: M_n(\mathbf{C}) \rightarrow \mathbf{C}$ is a polynomial function of matrix entries and therefore is continuous. The group $\mathrm{GL}_n(\mathbf{C})$ is the inverse image of \mathbf{C}^\times under \det , so it is the inverse image of an open set in \mathbf{C} under a continuous map. Thus $\mathrm{GL}_n(\mathbf{C})$ is open in $M_n(\mathbf{C})$. \square

Theorem 2.8. *The diagonalizable matrices in $M_n(\mathbf{C})$ contain a nonempty open subset of $M_n(\mathbf{C})$.*

Proof. Every matrix in $M_n(\mathbf{C})$ with distinct eigenvalues is diagonalizable. We will write down a specific diagonal matrix A and sketch two arguments explaining why there is a neighborhood of A in $M_n(\mathbf{C})$ containing only matrices with distinct eigenvalues, so some neighborhood of A in $M_n(\mathbf{C})$ is all diagonalizable matrices.

Consider the diagonal matrix

$$A = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n \end{pmatrix}$$

with diagonal entries $1, 2, \dots, n$. An $n \times n$ matrix that is close to A has eigenvalues that are close to $1, 2, \dots, n$. This is the idea of “continuous variation of roots.” Two $n \times n$ matrices that are close have characteristic polynomials whose coefficients are close, so their roots are close to each other when paired together in the right way.

If you don’t like the idea of continuous variation of roots, we can instead use the discriminant of a polynomial, which (like $b^2 - 4ac$ in the quadratic case) is a polynomial expression in the coefficients of the original polynomial and vanishes only when the polynomial has a repeated root. The characteristic polynomial of A is $(T - 1)(T - 2) \dots (T - n)$, which has nonzero discriminant since its roots are distinct. Matrices B near A have characteristic polynomials whose coefficients are close to the coefficients of the characteristic polynomial of A , so the discriminant of the characteristic polynomial of B is near the discriminant of the characteristic polynomial of A and thus is nonzero if B is sufficiently close to A . Therefore the characteristic polynomial of B has distinct roots, so B has distinct eigenvalues and is diagonalizable. \square

Remark 2.9. The diagonalizable matrices in $M_n(\mathbf{C})$ are not themselves an open set in $M_n(\mathbf{C})$ when $n \geq 2$. For instance, many nondiagonalizable matrices are very close to I_n :

$$\begin{pmatrix} 1 & \varepsilon & 0 & \cdots & 0 & 0 \\ 0 & 1 & \varepsilon & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \vdots & 1 & \varepsilon \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

For small ε this matrix is near the diagonalizable matrix I_n . The matrix has characteristic polynomial $(T - 1)^n$, so its only eigenvalue is 1. When $\varepsilon \neq 0$, the only eigenvectors of the matrix are scalar multiples of the first column, so the matrix is not diagonalizable when $n \geq 2$.

3. THE THEOREMS

Throughout this section, R denotes an arbitrary commutative ring. For $A \in M_n(R)$, its characteristic polynomial is $\chi_A(T) := \det(TI_n - A) \in R[T]$, whose coefficients are polynomials in the matrix entries of A .

Remark 3.1. Some texts define the characteristic polynomial of A as $\det(A - TI_n)$, rather than as $\det(TI_n - A)$. The two definitions differ only by an overall factor of $(-1)^n$ because $TI_n - A$ and $A - TI_n$ differ by an overall sign. Algebraically, it is better to use $\det(TI_n - A)$ since this has leading coefficient 1 rather than $(-1)^n$.

Our goal is to prove the following theorems about determinants by using polynomial identities over \mathbf{Z} and reduction to the complex case.

Theorem 3.2. For $A \in M_n(R)$ and $B \in M_m(R)$, let $A \oplus B := \begin{pmatrix} A & O \\ O & B \end{pmatrix}$, a block matrix in $M_{m+n}(R)$. Then $\det(A \oplus B) = \det(A) \det(B)$.

Theorem 3.2 can be proved by a direct calculation with matrices in R . We will prove it by reduction to the complex case simply as a warm-up to the more interesting identities below.

Theorem 3.3. For all A and B in $M_n(R)$, AB and BA have the same characteristic polynomial: $\det(TI_n - AB) = \det(TI_n - BA)$ in $R[T]$.

As a special case (looking at the constant terms of the characteristic polynomials), $\det(AB) = \det(BA)$, although this isn't really an honest consequence of Theorem 3.3 since the proof will use the fact that the product of two square matrices in either order has the same determinant.

Theorem 3.4 (Cayley-Hamilton). For $A \in M_n(R)$, $\chi_A(A) = O$ in $M_n(R)$.

Remark 3.5. If $A \in M_n(R)$ satisfies $\det(TI_n - A) = (T - \lambda_1) \cdots (T - \lambda_n)$ then for all $g(T) \in R[T]$ we have $\det(TI_n - g(A)) = (T - g(\lambda_1)) \cdots (T - g(\lambda_n))$. While this is a "general" kind of identity, it is not a universal algebraic identity because characteristic polynomials do not factor all the time. So we will not discuss a proof of such a formula using the method of universal identities.

To view the equation in Theorem 3.3 as a polynomial identity over \mathbf{Z} , we work in the ring

$$\mathbf{Z}[X_{11}, \dots, X_{nn}, Y_{11}, \dots, Y_{nn}, T].$$

The two matrices (X_{ij}) and (Y_{ij}) have entries in this ring and a special case of Theorem 3.3 says the two polynomials $\det(TI_n - (X_{ij})(Y_{ij}))$ and $\det(TI_n - (Y_{ij})(X_{ij}))$ are equal. Once we have such an equality of polynomials over \mathbf{Z} in $2n^2 + 1$ variables, we can specialize it to an identity in every commutative ring.

What about Theorem 3.4? For a "generic" matrix (X_{ij}) in n^2 indeterminates, write its characteristic polynomial as

$$(3.1) \quad \det(TI_n - (X_{ij})) = T^n + c_{n-1}(\underline{X})T^{n-1} + \cdots + c_1(\underline{X})T + c_0(\underline{X}),$$

where \underline{X} is an abbreviation for the listing of every X_{ij} and $c_k(\underline{X}) \in \mathbf{Z}[X_{11}, \dots, X_{nn}]$. A special case of Theorem 3.4 says that substituting the matrix (X_{ij}) for T on the right side of (3.1) yields the zero matrix:

$$(3.2) \quad (X_{ij})^n + c_{n-1}(\underline{X})(X_{ij})^{n-1} + \cdots + c_1(\underline{X})(X_{ij}) + c_0(\underline{X})I_n = O.$$

Equation (3.2) is not an equality of polynomials, but of matrices whose entries are polynomials (with integral coefficients). We need to think about (3.2) as a set of n^2 separate polynomial identities, one for each of the n^2 matrix entries on both sides. In each matrix entry the identity says a certain polynomial on the left side is 0. Once we know the polynomials in each matrix entry on the left are 0, (3.2) can be specialized to $M_n(R)$ for an

arbitrary commutative ring R by letting the X_{ij} 's be replaced by n^2 arbitrary elements of R .

Example 3.6. When $A = \begin{pmatrix} W & X \\ Y & Z \end{pmatrix}$,

$$\det(TI_2 - A) = \left| \begin{pmatrix} T - W & -X \\ -Y & T - Z \end{pmatrix} \right| = T^2 - (W + Z)T + (WZ - XY),$$

so the four polynomial identities connected to the Cayley-Hamilton theorem for 2×2 matrices come from the four matrix entries on both sides of the matrix identity

$$\begin{pmatrix} W & X \\ Y & Z \end{pmatrix}^2 - (W + Z) \begin{pmatrix} W & X \\ Y & Z \end{pmatrix} + (WZ - XY) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Since the three theorems in this section are stated over all commutative rings but can be viewed as consequences of identities of polynomials with integral coefficients, it suffices to verify these identities by viewing both sides as functions on a complex Euclidean space. The special feature of complex numbers we will exploit is that characteristic polynomials always factor completely: for $A \in M_n(\mathbf{C})$,

$$\chi_A(T) = \det(TI_n - A) = (T - \lambda_1)(T - \lambda_2) \cdots (T - \lambda_n),$$

for some $\lambda_i \in \mathbf{C}$.

4. PROOFS OF THEOREMS

To prove Theorems 3.2 and 3.4 we will use Theorem 2.8. To prove Theorem 3.3 we will use Theorem 2.7.

Proof. (of Theorem 3.2) View the equation $\det(A \oplus B) = \det(A) \det(B)$ as a polynomial identity in $n^2 + m^2$ variables. We want to prove it holds on an open subset of $M_n(\mathbf{C}) \times M_m(\mathbf{C})$. Pairs of matrices that are diagonalizable contain a nonempty open subset of $M_n(\mathbf{C}) \times M_m(\mathbf{C})$ by Theorem 2.8, so it suffices to prove the theorem when A and B are diagonalizable: each admits a basis of eigenvectors in \mathbf{C}^n and \mathbf{C}^m , respectively.

Let e_1, \dots, e_n be an eigenbasis for A and f_1, \dots, f_m be an eigenbasis for B : $Ae_i = \lambda_i e_i$ and $Bf_j = \mu_j f_j$. Then the sets $\{(e_i, 0)\}$ and $\{(0, f_j)\}$ are a basis of $\mathbf{C}^n \oplus \mathbf{C}^m$ that are eigenvectors for the matrix $A \oplus B$:

$$(A \oplus B)(e_i, 0) = (Ae_i, B(0)) = (\lambda_i e_i, 0) = \lambda_i (e_i, 0),$$

$$(A \oplus B)(0, f_j) = (A(0), Bf_j) = (0, \mu_j f_j) = \mu_j (0, f_j),$$

Since the determinant is the product of the eigenvalues for a basis of eigenvectors,

$$\det(A \oplus B) = \prod_i \lambda_i \cdot \prod_j \mu_j = \det(A) \det(B).$$

We are done by Theorem 2.6. □

Remark 4.1. Although Theorem 3.2 is about determinants of matrices over R , by replacing R with $R[T]$ we can turn it into a result about characteristic polynomials: $\chi_{A \oplus B}(T) = \chi_A(T) \chi_B(T)$. Indeed, since $TI_{m+n} - A \oplus B = (TI_n - A) \oplus (TI_m - B)$, we can use Theorem 3.2 with R replaced by $R[T]$, A replaced by $TI_n - A$, and B replaced by $TI_m - B$. (This identity of block matrices is also simple enough to be checked by a direct calculation over $R[T]$ without a reduction to the case of complex matrices.)

Proof. (of Theorem 3.3) We want to show

$$\det(tI_n - AB) = \det(tI_n - BA)$$

for complex $n \times n$ matrices A and B and complex numbers t . Specifically, the triples (A, B, t) fill out the space $M_n(\mathbf{C}) \times M_n(\mathbf{C}) \times \mathbf{C} \cong \mathbf{C}^{2n^2+1}$ and we need to prove the equality for a nonempty open set of such triples.

We will work with invertible A , *i.e.*, triples (A, B, t) in $\mathrm{GL}_n(\mathbf{C}) \times M_n(\mathbf{C}) \times \mathbf{C}$. Since $\mathrm{GL}_n(\mathbf{C})$ is open in $M_n(\mathbf{C})$ by Theorem 2.7, $\mathrm{GL}_n(\mathbf{C}) \times M_n(\mathbf{C}) \times \mathbf{C}$ is open in $M_n(\mathbf{C}) \times M_n(\mathbf{C}) \times \mathbf{C}$. When $A \in \mathrm{GL}_n(\mathbf{C})$, AB and BA are conjugate matrices: $AB = A(BA)A^{-1}$. Then $tI_n - AB$ and $tI_n - BA$ are also conjugate matrices:

$$A(tI_n - BA)A^{-1} = A(tI_n)A^{-1} - A(BA)A^{-1} = tI_n - AB,$$

where we commuted the scalar diagonal matrix tI_n past A so A and A^{-1} cancel. Conjugate matrices have the same determinant, so $\det(tI_n - AB) = \det(tI_n - BA)$. By Theorem 2.6 we are done. \square

Here is a generalization of Theorem 3.3 to rectangular matrices: when $A \in M_{m \times n}(R)$ and $B \in M_{n \times m}(R)$, $T^n \det(TI_m - AB) = T^m \det(TI_n - BA)$. (Since $AB \in M_m(R)$ and $BA \in M_n(R)$, these determinants make sense.) Do you see a way to prove this by reduction to the complex case, or perhaps turning it into a special case of Theorem 3.3 using characteristic polynomials of $(m+n) \times (m+n)$ matrices? A short proof that bypasses reduction to the complex case runs as follows [2]. Set $M = \begin{pmatrix} TI_m & A \\ B & I_n \end{pmatrix}$ and $N = \begin{pmatrix} I_m & O \\ -B & TI_n \end{pmatrix}$. Then $MN = \begin{pmatrix} TI_m - AB & TA \\ O & TI_n \end{pmatrix}$ and $NM = \begin{pmatrix} TI_m & A \\ O & TI_n - BA \end{pmatrix}$. Equate $\det(MN)$ and $\det(NM)$.

Proof. (of Theorem 3.4) We want to show each matrix entry on the left side of (3.2) is the polynomial 0. It suffices to check such vanishing when $A \in M_n(\mathbf{C})$ is diagonalizable, since such matrices contain a nonempty open subset of $M_n(\mathbf{C})$ by Theorem 2.8. Let $U \in \mathrm{GL}_n(\mathbf{C})$ be a matrix conjugating A into a diagonal matrix: $D := UAU^{-1}$ is diagonal. Then $D^k = UA^kU^{-1}$ for all $k \geq 0$. Since A and D are conjugate, $\chi_A(T) = \chi_D(T)$ in $\mathbf{C}[T]$. Writing the common characteristic polynomial of A and D as $T^n + c_{n-1}T^{n-1} + \cdots + c_1T + c_0$, we have

$$\begin{aligned} \chi_D(D) &= D^n + c_{n-1}D^{n-1} + \cdots + c_1D + c_0I_n \\ &= UA^nU^{-1} + c_{n-1}UA^{n-1}U^{-1} + \cdots + c_1UAU^{-1} + c_0I_n \\ &= U(A^n + c_{n-1}A^{n-1} + \cdots + c_1A + c_0I_n)U^{-1} \\ &= U\chi_D(A)U^{-1} \\ &= U\chi_A(A)U^{-1}. \end{aligned}$$

Thus $\chi_A(A) = U^{-1}\chi_D(D)U$, so it suffices to check $\chi_D(D) = O$.

Let D have diagonal entries $\lambda_1, \dots, \lambda_n$, so

$$(4.1) \quad T^n + c_{n-1}T^{n-1} + \cdots + c_1T + c_0 = \prod_{i=1}^n (T - \lambda_i).$$

Substitution of a single matrix into a polynomial identity in $\mathbf{C}[T]$ is a homomorphism (Remark 2.4), so (4.1) at $T = D$ becomes

$$\begin{aligned} D^n + c_{n-1}D^{n-1} + \cdots + c_1D + c_0I_n &= \prod_{i=1}^n (D - \lambda_i I_n) \\ &= \prod_{i=1}^n \begin{pmatrix} \lambda_1 - \lambda_i & 0 & \cdots & 0 \\ 0 & \lambda_2 - \lambda_i & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n - \lambda_i \end{pmatrix}. \end{aligned}$$

The i -th entry of the i -th matrix is 0, so the product of these diagonal matrices is the zero matrix. Therefore $\chi_D(D) = O$. \square

It is worth reviewing the reduction steps in this proof of the Cayley-Hamilton theorem. We start with a single matrix identity (the Cayley-Hamilton theorem) that is to be proved in every commutative ring. Rather than actually working in a general commutative ring, we recognize the universal nature of the identity: it is (as a special case) an identity of matrices whose entries are polynomials in many variables with integer coefficients, so it is a set of many polynomial identities (one for each matrix entry). If we can prove all of these polynomial identities then we can specialize them into every commutative ring. To prove these polynomial identities, we treat the Cayley-Hamilton theorem as an identity of complex matrices, where we only have to verify it on diagonalizable matrices by Theorems 2.6 and 2.8. The Cayley-Hamilton theorem is insensitive to matrix conjugation, so we can reduce further from diagonalizable matrices to diagonal matrices, where the theorem is a simple direct computation.

There are other proofs of the Cayley-Hamilton theorem, *e.g.*, using rational or Jordan canonical form. By using the viewpoint of (universal) polynomial identities we have proved the Cayley-Hamilton theorem in one stroke in all commutative rings (not just fields) and the only matrix calculation we made used multiplication of diagonal matrices, which is very easy.

5. CONSEQUENCE OF THE CAYLEY-HAMILTON THEOREM

In this section we put the Cayley-Hamilton theorem to work. Throughout, R is a commutative ring. Everything is based on the next lemma, which says each square matrix A has an “almost inverse”: a matrix C such that $AC = CA$ and this is a special scalar diagonal matrix.

Lemma 5.1. *For $A \in M_n(R)$, there is a matrix $C \in M_n(R)$ such that $AC = CA = (\det A)I_n$.*

Proof. Let A have characteristic polynomial $\det(TI_n - A) = T^n + c_{n-1}T^{n-1} + \cdots + c_1T + c_0$, so (setting $T = 0$) $c_0 = \det(-A) = \pm \det A$. From the Cayley-Hamilton theorem,

$$A^n + c_{n-1}A^{n-1} + \cdots + c_1A + c_0I_n = O.$$

Bring c_0I_n to the right side and factor A from all other terms:

$$A(A^{n-1} + c_{n-1}A^{n-2} + \cdots + c_1I_n) = -c_0I_n = \pm(\det A)I_n.$$

Thus $AC = (\det A)I_n$, where $C = \pm(A^{n-1} + c_{n-1}A^{n-2} + \cdots + c_1I_n)$. Since C is a polynomial in A , C and A commute, so $CA = (\det A)I_n$ too. \square

We chose the letter C because it is called the cofactor matrix of A . Working in the ring $\mathbf{Z}[X_{11}, \dots, X_{nn}]$ and taking $A = (X_{ij})$, the proof of Lemma 5.1 shows the cofactor matrix of (X_{ij}) has entries given by universal polynomials in the matrix entries of A . Explicitly, the (i, j) entry of C is $(-1)^{i+j}A_{ji}$, where A_{ji} is the determinant of the matrix obtained from A by removing the j th row and i th column. For instance, in the 2×2 case, if $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ then $C = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$ and $AC = CA = \begin{pmatrix} ad-bc & 0 \\ 0 & ad-bc \end{pmatrix}$. However, we did not need to know this explicit formula for the entries of C to prove Lemma 5.1.

Theorem 5.2. *For $A \in M_n(R)$, $A \in \text{GL}_n(R)$ if and only if $\det A \in R^\times$, in which case $A^{-1} = \frac{1}{\det A}C$.*

Proof. If $A \in \text{GL}_n(R)$, let B be the inverse matrix, so $AB = I_n$. Taking determinants, $(\det A)(\det B) = 1$ in R , so $\det A \in R^\times$. Conversely, suppose $\det A \in R^\times$. By Lemma 5.1 we have the formula $AC = CA = (\det A)I_n$. Since $\det A$ is invertible, the matrix $\frac{1}{\det A}C$ is a multiplicative inverse for A , so $A \in \text{GL}_n(R)$. \square

The formula $A^{-1} = \frac{1}{\det A}C$ shows that we only have to do one division to invert a matrix: divide by $\det A$. To compute the inverse matrix, all other calculations are additions and multiplications since C is a polynomial in A .

The following interesting application of Lemma 5.1 may come as a surprise. Notice the module that occurs in the theorem need not be free.

Theorem 5.3. *Let M be a finitely generated R -module. If $\varphi: M \rightarrow M$ is R -linear and surjective then it is an isomorphism.*

Proof. We want to show φ is injective. We will view M as an $R[T]$ -module by letting T act through φ : $Tm = \varphi(m)$ for $m \in M$. More generally, polynomials $f(T) \in R[T]$ act on M by $f(T)m = f(\varphi)(m)$: $(c_i T^i)(m) = c_i \varphi^i(m)$. Here φ^i means the i -fold composite of φ with itself: $\varphi^2(m) = \varphi(\varphi(m))$, and so on. Let x_1, \dots, x_n be a spanning set for M as an R -module: $M = \sum_{i=1}^n R x_i$. Because $\varphi: M \rightarrow M$ is onto we can write $x_i = \varphi(y_i)$ for some $y_i \in M$. Write $y_i = \sum_{j=1}^n a_{ij} x_j$, with $a_{ij} \in R$, so for $i = 1, 2, \dots, n$,

$$(5.1) \quad x_i = \varphi(y_i) = \sum_{j=1}^n a_{ij} \varphi(x_j) = \sum_{j=1}^n a_{ij} T \cdot x_j.$$

Since M is an $R[T]$ -module, the matrix ring $M_n(R[T])$ acts on the n -tuples M^n from the left. This lets us write (5.1) for $i = 1, \dots, n$ as a single matrix equation:

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_{11}T & \cdots & a_{1n}T \\ \vdots & \ddots & \vdots \\ a_{n1}T & \cdots & a_{nn}T \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

Bring all terms to the left side:

$$(I_n - TA) \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix},$$

where $A = (a_{ij}) \in M_n(R)$. Multiplying both sides on the left by the cofactor matrix of $I_n - TA$ in $M_n(R[T])$, from Lemma 5.1 we get

$$\begin{pmatrix} d(T) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & d(T) \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix},$$

where $d(T) = \det(I_n - TA) \in R[T]$. Therefore $d(T)x_i = 0$ for all i . Since the x_i 's span M , $d(T)m = 0$ for all $m \in M$. The polynomial $d(T)$ has constant term $d(0) = \det(I_n) = 1$, so $d(T) = 1 + f(T)T$ for some $f(T) \in R[T]$. Since T acts on M as φ , for all $m \in M$

$$0 = d(T)m = (1 + f(T)T)(m) = m + f(\varphi)(\varphi(m)).$$

In particular, if $\varphi(m) = 0$ then $0 = m + f(\varphi)(0) = m$, so $\ker \varphi = \{0\}$. \square

Since $d(T)m = 0$ for all $m \in M$, M is a torsion module over $R[T]$. This is like the setting of a finite-dimensional vector space over a field F , which is a torsion module over $F[T]$ when T acts on V by some linear operator $\varphi: V \rightarrow V$. It's again worth stressing M only has to be a finitely generated R -module, not a finite free R -module.

Remark 5.4. Theorem 5.3 is due independently to Strooker [3] and Vasconcelos [4]. For some related results, see [1] and [5].

Corollary 5.5. *Let M be a finite free R -module of rank $n \geq 1$. Every spanning set of M of size n is a basis of M .*

Proof. Let x_1, \dots, x_n be a spanning set of size n . We want to show x_1, \dots, x_n is a linearly independent set. Pick a basis e_1, \dots, e_n of M . (There is a basis since M is finite free over R .) Define a linear map $\varphi: M \rightarrow M$ by $\varphi(e_i) = x_i$, i.e., $\varphi(\sum a_i e_i) = \sum a_i x_i$. Since the x_i 's span M , φ is onto. Hence φ is an isomorphism by Theorem 5.3, so linear independence of the set $\{e_i\}$ carries over to the set $\{\varphi(e_i)\} = \{x_i\}$. \square

We showed a spanning set of the right size in a finite free R -module is linearly independent, but it is false that a linearly independent set of the right size is a spanning set (in general). Consider the vectors $2e_1, \dots, 2e_n$ in \mathbf{Z}^n .

Corollary 5.6. *Let M and N be isomorphic finitely generated R -modules. Every R -linear surjection from M to N is an isomorphism.*

Proof. There is some isomorphism $M \rightarrow N$ by hypothesis. Call it f . If $\varphi: M \rightarrow N$ is a linear surjection then $f^{-1} \circ \varphi: M \rightarrow M$ is a linear surjection of M to itself, so $f^{-1} \circ \varphi$ is an isomorphism by Theorem 5.3. Composing this with f (an isomorphism) shows $f \circ f^{-1} \circ \varphi = \varphi$ is an isomorphism. \square

Corollary 5.7. *Let A be an R -algebra with identity, possibly noncommutative, which is finitely generated as an R -module. If $x, y \in A$ satisfy $xy = 1$ then $yx = 1$.*

Proof. Let $f: A \rightarrow A$ by $f(a) = xa$. Then f is R -linear. Since $f(y) = xy = 1$, f is onto: $f(ya) = xya = a$. By Theorem 5.3, f is one-to-one as well. Then since $f(yx) = xyx = x$ and $f(1) = x1 = x$ we get $yx = 1$. \square

Example 5.8. Let $A = M_n(R)$. This R -algebra is a finitely generated R -module, since a generating set is the matrices with 1 in one component and 0 elsewhere. Suppose two matrices M and N in $M_n(R)$ satisfy $MN = I_n$. Then $NM = I_n$, so M and N are 2-sided inverses of each other.

The next result does not use the ideas we have mentioned so far, but it is close in spirit to Theorem 5.3 (with rings in place of modules and ring homomorphisms in place of linear maps).

Theorem 5.9. *Let R be a Noetherian ring. If $\varphi: R \rightarrow R$ is a surjective ring homomorphism then φ is an isomorphism.*

A ring is called *Noetherian* when all of its ideals are finitely generated. Most rings that arise naturally in algebra are Noetherian. The Noetherian condition in Theorem 5.9 is a replacement in the ring setting for the finite generatedness of modules in Theorem 5.3.

When a Noetherian ring contains an increasing chain of ideals

$$I_1 \subset I_2 \subset I_3 \subset \cdots$$

then the chain must stabilize: $I_n = I_{n+1} = I_{n+2} = \cdots$ for some n . Indeed, the union $\bigcup_{k \geq 1} I_k$ is an ideal (why?) so by hypothesis it is finitely generated. Those generating elements each lie in some I_k , so all of them lie in a common I_n because the chain is increasing. That means the whole union is I_n , so the chain stabilizes from the n th ideal onwards. We will use this in the proof of Theorem 5.9.

Proof. We want to show φ is injective. Every iterate φ^n is a surjective ring homomorphism. Let $K_n = \ker(\varphi^n)$, so every K_n is an ideal in R . These ideals form an increasing chain:

$$K_1 \subset K_2 \subset \cdots \subset K_n \subset \cdots .$$

Since R is Noetherian (that is, all ideals in R are finitely generated) this chain stabilizes. In particular, $K_n = K_{n+1}$ for some n . The inclusion $K_{n+1} \subset K_n$ tells us that if $\varphi^{n+1}(x) = 0$ for some $x \in R$ then $\varphi^n(x) = 0$. (This is only for one n ; we can't “induct” down on n to get $n = 1$.)

Assume $y \in \ker(\varphi)$. We want to show $y = 0$. Since φ^n is onto, we can write $y = \varphi^n(x)$ for some $x \in R$. Thus $\varphi(y) = \varphi^{n+1}(x)$. Since $\varphi(y) = 0$ we conclude $\varphi^n(x) = 0$ by the previous paragraph. That means $y = 0$. \square

REFERENCES

- [1] M. Orzech, “Onto Endomorphisms are Isomorphisms,” *Amer. Math. Monthly* **78** (1971), 357–362.
- [2] J. Schmid, “A Remark on Characteristic Polynomials,” *Amer. Math. Monthly* **77** (1970), 998–999.
- [3] J. R. Strooker, “Lifting Projectives,” *Nagoya Math. J.* **27** (1966), 747–751.
- [4] W. Vasconcelos, “On Finitely Generated Flat Modules,” *Trans. Amer. Math. Soc.* **138** (1969), 505–512.
- [5] W. Vasconcelos, “Injective Endomorphisms of Finitely Generated Modules,” *Proc. Amer. Math. Soc.* **25** (1970), 900–901.