UNIVERSAL IDENTITIES

KEITH CONRAD

1. Introduction

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

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

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

A direct calculation shows that (1.1) holds in any commutative ring. That is, (1.1) is true when the 4 parameters are elements of any 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

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

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 over all commutative rings is multiplicativity of determinants. In the 2×2 case this says

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

or equivalently

$$(ad-bc)(a'd'-b'c') = (aa'+bc')(cb'+dd') - (ab'+bd')(ca'+dc'). \tag{1.3}$$

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 over all commutative rings can be proved by working only over $\mathbf{C}$. This is a really significant idea! In Section 3 we'll state two 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 degrees of f in each of $X_1, \dots, X_n$ often won't matter. We also might write $f(X_1, \dots, X_n)$ simply as $f(\underline{X})$.

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

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

2. Reduction to the Complex Case

An identity over 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 any 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 $\mathbf{Z}[X_1, \dots, X_n] \to R$.

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

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

The catchphrase for this idea is "substitution (or specialization) is a ring homomorphism."

Example 2.2. Substituting into both sides of the polynomial identity (1.2) elements of any commutative ring R 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 $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ and Y with $\left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$ then $X^2 - Y^2$ goes to $\left(\begin{smallmatrix} 0 & 2 \\ -2 & 0 \end{smallmatrix}\right)$ while $(X+Y)(X-Y)$ goes to $\left(\begin{smallmatrix} -1 & 2 \\ -2 & 1 \end{smallmatrix}\right)$, 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 $f(\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right), \left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right))$ 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 $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$. 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 any constant terms 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 any $n \times n$ 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 any 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 any elements of any commutative ring: a polynomial identity with integral coefficients remains true under specialization into any 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

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

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

To prove an identity over 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 which are used to verify (1.2) are the same as the ones used to verify (1.1), so there is no non-trivial 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 (non-algebraic) topological hypothesis which occurs.

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

Let $U \subset \mathbf{C}^{n+1}$ be a nonempty open set in $\mathbf{C}^{n+1}$ where f vanishes: if $(z_1, \ldots, z_{n+1}) \in U$ then $f(z_1, \ldots, z_{n+1}) = 0$. From the topology of $\mathbf{C}^{n+1}$, U contains (around any 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 any point $(z_1, \ldots, z_{n+1}) \in U_1 \times \cdots \times U_{n+1}$ and consider the *one-variable* polynomial derived from $f(X_1, \ldots, X_{n+1})$

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

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

$$c_i(z_1, \ldots, z_n) = 0 \tag{2.2}$$

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

Theorem 2.6 goes through with $\mathbf{R}$ in place of $\mathbf{C}$ by 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 over $\mathbf{C}$, so we only stated the theorem that way. However, working over $\mathbf{R}$ 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 subset 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 subset of $\mathbf{R}^2$, then $f(X, Y) = 0$. This reasoning extends to more than 2 variables and complex coefficients if the reader has any experience with differential or algebraic geometry.

Combining Theorems 2.1 and 2.6, we have the following procedure for reducing the verification of a polynomial identity over all commutative rings (such as (1.2)) to its verification as a polynomial identity over $\mathbf{C}$:

- Express the identity as $f(a_1, \ldots, a_n) = g(a_1, \ldots, a_n)$ where f and g are in some $\mathbf{Z}[X_1, \ldots, X_n]$ and $a_1, \ldots, a_n$ run over elements of any commutative ring.
- Verify $f(z_1, \ldots, z_n) = g(z_1, \ldots, z_n)$ as $(z_1, \ldots, z_n)$ runs over an open subset of $\mathbf{C}^n$.
- By Theorem 2.6, $f(X_1, \ldots, X_n) = g(X_1, \ldots, X_n)$ in $\mathbf{Z}[X_1, \ldots, X_n]$.
- By Theorem 2.1, $f(a_1, \ldots, a_n) = g(a_1, \ldots, a_n)$ where the a_i's are taken from any commutative ring. That is, our identity is true over 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 $\mathrm{M}_n(\mathbf{C})$, which can be thought of as $\mathbf{C}^{n^2}$ in a natural way. The topology of $\mathrm{M}_n(\mathbf{C})$ 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 $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ is all matrices $\left(\begin{smallmatrix} a+\varepsilon_1 & b+\varepsilon_2 \\ c+\varepsilon_3 & d+\varepsilon_4 \end{smallmatrix}\right)$ 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 $\mathrm{M}_n(\mathbf{C})$. 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* $\mathrm{M}_n(\mathbf{C})$*.*

Proof. The determinant function $\det\colon \mathrm{M}_n(\mathbf{C}) \to \mathbf{C}$ 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 under a continuous map. Therefore it is open. □

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

Proof. Since a matrix with distinct eigenvalues is diagonalizable, we will write down a matrix A with distinct eigenvalues and sketch two arguments why it has a neighborhood in $\mathrm{M}_n(\mathbf{C})$ of matrices with distinct eigenvalues, so a small neighborhood of A in $\mathrm{M}_n(\mathbf{C})$ 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 entries $1, 2, \ldots, n$. Any $n \times n$ matrix which is close to A has eigenvalues that are close to $1, 2, \ldots, n$. 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 $(T-1)(T-2)\cdots(T-n)$, 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. □

Remark 2.9. The diagonalizable matrices in $\mathrm{M}_n(\mathbf{C})$ are not themselves an open subset when $n \geq 2$. For instance, we can find many nondiagonalizable matrices very close to I_n:

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

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 can't be diagonalized when $n \geq 2$.

3. The Theorems

Throughout this section, R denotes an arbitrary commutative ring. For any matrix $A \in \mathrm{M}_n(R)$, 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 makes the leading coefficient 1 rather than some screwy sign.

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 $A \in \mathrm{M}_n(R)$ and $B \in \mathrm{M}_m(R)$, let $A \oplus B := \left(\begin{smallmatrix} A & O \\ O & B \end{smallmatrix}\right)$, a block matrix in $\mathrm{M}_{m+n}(R)$. 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 any A and B in $\mathrm{M}_n(R)$, 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 $A \in \mathrm{M}_n(R)$, $\chi_A(A) = O$ in $\mathrm{M}_n(R)$.*

Remark 3.5. If $A \in \mathrm{M}_n(R)$ satisfies $\det(TI_n - A) = (T - \lambda_1)\cdots(T - \lambda_n)$ then for any $g(T) \in R[T]$ 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

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

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 equality of polynomials over $\mathbf{Z}$ in $2n^2 + 1$ variables, we can specialize it to an identity in any commutative ring.

What about Theorem 3.4? Write

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

where $c_k(\underline{X}) \in \mathbf{Z}[X_{11}, \ldots, X_{nn}]$. 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:

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

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. In each 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 $\mathrm{M}_n(R)$ for any commutative ring R by letting the X_{ij}'s be replaced by any n^2 elements of R.

Example 3.6. When $A = \left(\begin{smallmatrix} W & X \\ Y & Z \end{smallmatrix}\right)$, $\det(TI_2 - A) = \left|\begin{smallmatrix} T-W & -X \\ -Y & T-Z \end{smallmatrix}\right| = T^2 - (W+Z)T + (WZ - XY)$, so the 4 polynomial identities connected to the Cayley-Hamilton theorem for 2 × 2 matrices come from the 4 matrix entries on both sides of the 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 both of the theorems are stated over any commutative ring 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 complex Euclidean space (of matrices).

The special feature of complex numbers we will exploit is that characteristic polynomials always factor completely: for $A \in \mathrm{M}_n(\mathbf{C})$,

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

where $\lambda_i \in \mathbf{C}$.

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 $\mathrm{M}_n(\mathbf{C}) \times \mathrm{M}_m(\mathbf{C})$. Pairs of matrices which are diagonalizable contain an open subset of $\mathrm{M}_n(\mathbf{C}) \times \mathrm{M}_m(\mathbf{C})$ 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 $e_1, \ldots, e_n$ be an eigenbasis for A and $f_1, \ldots, f_m$ 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 $\mathbf{C}^n \oplus \mathbf{C}^m$ which 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 (with multiplicity),

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

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 $\mathrm{M}_n(\mathbf{C}) \times \mathrm{M}_n(\mathbf{C}) \times \mathbf{C} \cong \mathbf{C}^{2n^2+1}$ and we need to prove the equality for an open set of such triples.

We will work with invertible A, *i.e.*, triples (A, B, t) in $\mathrm{GL}_n(\mathbf{C}) \times \mathrm{M}_n(\mathbf{C}) \times \mathbf{C}$. Since $\mathrm{GL}_n(\mathbf{C})$ is open in $\mathrm{M}_n(\mathbf{C})$ by Theorem 2.7, $\mathrm{GL}_n(\mathbf{C}) \times \mathrm{M}_n(\mathbf{C}) \times \mathbf{C}$ is open in $\mathrm{M}_n(\mathbf{C}) \times \mathrm{M}_n(\mathbf{C}) \times \mathbf{C}$. 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. □

Here is a generalization of Theorem 3.3 to rectangular matrices: when $A \in \mathrm{M}_{m\times n}(R)$ and $B \in \mathrm{M}_{n\times m}(R)$, $T^n \det(TI_m - AB) = T^m \det(TI_n - BA)$. (Since $AB \in \mathrm{M}_m(R)$ and $BA \in \mathrm{M}_n(R)$, 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 $M = \left(\begin{smallmatrix} TI_m & A \\ B & I_n \end{smallmatrix}\right)$ and $N = \left(\begin{smallmatrix} I_m & O \\ -B & TI_n \end{smallmatrix}\right)$. Then $MN = \left(\begin{smallmatrix} TI_m - AB & TA \\ O & TI_n \end{smallmatrix}\right)$ and $NM = \left(\begin{smallmatrix} TI_m & A \\ O & TI_n - AB \end{smallmatrix}\right)$. 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 $A \in \mathrm{M}_n(\mathbf{C})$ is diagonalizable, since such matrices contain an open subset of $\mathrm{M}_n(\mathbf{C})$ 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 any $k \geq 0$. 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 it suffices to check $\chi_D(D) = O$.

Let D have diagonal entries $\lambda_1, \ldots, \lambda_n$, so its characteristic polynomial is $\chi_D(T) = \prod_{i=1}^n (T - \lambda_i)$. Then

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

The i-th entry of the i-th matrix is 0, so multiplying them all together produces the zero matrix. □

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) which is to be proved over 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 any 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 even further to the case of diagonal matrices.

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 over all commutative rings (not just over fields) and the only matrix calculation we made used multiplication of diagonal matrices, which is very easy.

5. Consequences of the Cayley-Hamilton theorem

We now put the Cayley-Hamilton theorem to work. Everything will follow from the next lemma, which says every square matrix has an "almost inverse": a matrix whose product in either order is a scalar diagonal matrix.

Lemma 5.1. *Let R be a commutative ring and $A \in \mathrm{M}_n(R)$. There is a matrix $C \in \mathrm{M}_n(R)$ 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. □

We chose the letter C because it is called the cofactor matrix of A. Working in the ring $\mathbf{Z}[X_{11}, \ldots, X_{nn}]$ and taking $A = (X_{ij})$, the proof of Lemma 5.1 shows the cofactor matrix has entries given by universal polynomial functions 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 jth row and ith column. (For instance, in the 2×2 case if $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ then $C = \left(\begin{smallmatrix} d & -b \\ -c & a \end{smallmatrix}\right)$ and $AC = CA = \left(\begin{smallmatrix} ad-bc & 0 \\ 0 & ad-bc \end{smallmatrix}\right)$.) However, we did *not* need to know this explicit formula for the entries of C to prove Lemma 5.1.

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

Proof. If $A \in \mathrm{GL}_n(R)$, 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 $A \in \mathrm{GL}_n(R)$. □

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 $\varphi\colon M \to M$ 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: $T \cdot m = \varphi(m)$ for $m \in M$. More generally, any polynomial $f(T) \in R[T]$ acts on M by $f(T) \cdot m = f(\varphi)(m)$: $(\sum c_i T^i)(m) = \sum c_i \varphi^i(m)$. Here φ^i means the i-fold composite of φ with itself: $\varphi^2(m) = \varphi(\varphi(m))$, and so on.

Let $x_1, \ldots, x_n$ be a spanning set for M as an R-module: $M = \sum_{i=1}^n Rx_i$. Because $\varphi\colon M \to M$ 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 $i = 1, 2, \ldots, n$,

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

Since M is an $R[T]$-module, the matrix ring $\mathrm{M}_n(R[T])$ acts on the n-tuples M^n from the left. This lets us write (5.1) over $1 \le i \le n$ as a 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 $A = (a_{ij}) \in \mathrm{M}_n(R)$. Multiplying both sides on the left by the cofactor matrix of $I_n - TA$ in $\mathrm{M}_n(R[T])$, from Lemma 5.1, gives

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

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 through 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 $0 = m + f(\varphi)(0) = m$. □

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 $\varphi\colon V \to V$. It's again worth stressing M only has to be finitely generated, not finite free.

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 $n \geq 1$. Any spanning set of size n is a basis.*

Proof. Let $x_1, \ldots, x_n$ be a spanning set of size n. We want to show $x_1, \ldots, x_n$ is a linearly independent set. Pick a basis $e_1, \ldots, e_n$ of M. (There is a basis since M is finite free over R.) Define a linear map $\varphi\colon M \to M$ 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 $\{\varphi(e_i)\} = \{x_i\}$. □

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 $2e_1, \ldots, 2e_n$ in $\mathbf{Z}^n$.

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

Proof. There is some isomorphism $M \to N$ by hypothesis. Call it f. If $\varphi\colon M \to N$ is any linear surjection then $f^{-1} \circ \varphi\colon M \to M$ 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. □

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 $f\colon A \to A$ 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 $f(1) = x \cdot 1 = x$ we get $yx = 1$. □

Example 5.8. As a review of the proof, take $A = \mathrm{M}_n(R)$. Suppose two matrices M and N in $\mathrm{M}_n(R)$ satisfy $MN = I_n$. Then $M(Nv) = v$ for all $v \in R^n$, so $M\colon R^n \to R^n$ is onto, and therefore M is one-to-one since R^n is finitely generated. So since $M(NM(v)) = (MN)(Mv) = Mv$, we have $NM(v) = v$ for all v. Thus $NM = I_n$. So a one-sided inverse for a square matrix over a commutative ring is in fact a 2-sided inverse (which, by Lemma 5.1, is even a polynomial in the matrix).

The next result does not use any of 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 $\varphi\colon R \to R$ 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 on a commutative ring is the ring-theoretic analogue of finite generatedness of a module. 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 $\cup_{k\geq 1} I_k$ 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 nth 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 $y = 0$. □

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.