$$= k_{33} + k_{22} + k_{11}$$

$$C_3 = \sum_{i_1 < i_2 < i_3} \frac{\partial^3}{\partial k_{i_1 i_1} \partial k_{i_2 i_2} \partial k_{i_3 i_3}} D = \frac{\partial^3 D}{\partial k_{11} \partial k_{22} \partial k_{33}} = \begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = 1. \tag{9}$$

Thus, our desired polynomial $g(t)$ is given by

$$g(t) = C_0 + C_1 t + C_2 t^2 + C_3 t^3, \tag{10}$$

where C_0, C_1, C_2, and C_3 are given by (6)–(9), respectively.

Taking

$$\underline{K} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 2 & 2 \\ 3 & 2 & 3 \end{pmatrix},$$

we have:

$$C_0 = \begin{vmatrix} 1 & 2 & 3 \\ 4 & 2 & 2 \\ 3 & 2 & 3 \end{vmatrix} = 4, \qquad C_1 = \begin{vmatrix} 2 & 2 \\ 2 & 3 \end{vmatrix} + \begin{vmatrix} 1 & 3 \\ 3 & 3 \end{vmatrix} + \begin{vmatrix} 1 & 2 \\ 4 & 2 \end{vmatrix} = 2 - 6 - 6 = -10,$$

$$C_2 = 3 + 2 + 1 = 6, \qquad C_3 = 1.$$

Substituting these values in (10), we obtain

$$g(t) = -4 - 10t + 6t^2 + t^3.$$

REMARK. The relationship between $g(t)$ of (3) and $P(\lambda)$ of (1) is given by $g(-\lambda) = P(\lambda)$, where going from (1) to (3), we have put $\underline{K} = \underline{A}$ and $t = -\lambda$.

Reference

[1] L. A. Zadeh and C. A. Desoer, Linear System Theory: The State Approach, McGraw-Hill, New York, 1963, pp. 303–305.

Matrices as Sums of Invertible Matrices

N. J. LORD
Tonbridge School
Kent TN9 1JP England

While it is a trivial truism that not every matrix is invertible, it does not seem to be well known that every matrix can be expressed as the sum of two invertible matrices. The proof of this makes a good exercise in elementary linear algebra and, although a direct proof is short, we have expanded the discussion to indicate several contrasting lines of attack.

For convenience we shall adopt the following notation:

$\mathbb{F}$ will denote the field under consideration;

q will denote the number of elements of $\mathbb{F}$ if $\mathbb{F}$ is finite;

$M(n, \mathbb{F})$ will denote the ring of $n \times n$ matrices with entries in $\mathbb{F}$, where $n > 1$;

$G(n, \mathbb{F})$ will denote the group of invertible matrices in $M(n, \mathbb{F})$;

I (or I_n, for emphasis) will denote the $n \times n$ identity matrix.

The theorem that we are going to prove then is as follows:

THEOREM. *Let A be in $M(n,\mathbb{F})$. Then A can be expressed as the sum of two elements of $G(n,\mathbb{F})$ where*

(i) *the summands may be taken as distinct unless A is the zero matrix and $\mathbb{F}$ has characteristic 2;*

(ii) *the decomposition is unique only if A is a nonzero 2×2 matrix with entries in the field with two elements.*

We deal first with the easiest case: $\mathbb{F}$ is infinite. For x in $\mathbb{F}$, let $p(x)$ denote the nonzero polynomial function $\det(A-xI)$. Since p has degree n, $p(x)$ vanishes for at most n values of x, so we can certainly find x_0 in $\mathbb{F}$ with $x_0\neq 0$, $p(x_0)\neq 0$ and $A\neq 2x_0I$.

Then $x_0I\in G(n,\mathbb{F})$, because $x_0\neq 0$; $A-x_0I\in G(n,\mathbb{F})$, because $p(x_0)\neq 0$; and $x_0I\neq A-x_0I$, because $A\neq 2x_0I$. Thus $A=(A-x_0I)+x_0I$ provides a splitting of A as the sum of two distinct elements of $G(n,\mathbb{F})$. (This argument will also carry through for a finite field provided that $n+3\leqslant q$.)

For $\mathbb{F}$ finite and $q>2$, we can use a counting argument based on a comparison of the sizes of $M(n,\mathbb{F})$ and $G(n,\mathbb{F})$. First note that an element of $M(n,\mathbb{F})$ can be obtained by putting any of the q elements of $\mathbb{F}$ into n^2 'slots' so $|M(n,\mathbb{F})|$, the cardinality of $M(n,\mathbb{F})$, is q^{n^2}.

Next, an element of $G(n,\mathbb{F})$ is characterised by the fact that its columns form an ordered linearly independent set. The first column is subject only to the restriction that it is nonzero: q^n-1 choices. The second column is subject only to the restriction that it is not linearly dependent on the first column: q^n-q choices. Continuing in this manner we see that

$$\begin{aligned}|G(n,\mathbb{F})| &= (q^n-1)(q^n-q)(q^n-q^2)\cdots(q^n-q^{n-1})\\ &= q^{\frac{1}{2}n(n-1)}(q-1)(q^2-1)\cdots(q^n-1).\end{aligned}$$

Further progress hinges on establishing the inequality:

$$2|G(n,\mathbb{F})|>|M(n,\mathbb{F})|+1.$$

In view of our formulae, this is equivalent to showing

$$2(q-1)(q^2-1)\cdots(q^n-1)>q^{\frac{1}{2}n(n+1)}+q^{-\frac{1}{2}n(n-1)}$$

or

$$(1-q^{-1})(1-q^{-2})\cdots(1-q^{-n})>\tfrac{1}{2}(1+q^{-n^2}).$$

Since $q\geqslant 3$,

$$(1-q^{-1})(1-q^{-2})\cdots(1-q^{-n})\geqslant(1-3^{-1})(1-3^{-2})\cdots(1-3^{-n})$$

and

$$\tfrac{1}{2}(1+q^{-n^2})\leqslant\tfrac{1}{2}(1+3^{-n^2})<\tfrac{1}{2}(1+3^{-n}).$$

So it is enough to show that $(1-3^{-1})(1-3^{-2})\cdots(1-3^{-n})>\frac{1}{2}(1+3^{-n})$, which is easily done by induction on n $(n\geqslant 2)$.

Next, define a map T_A from $G(n,\mathbb{F})$ to $M(n,\mathbb{F})$ by $T_A(X)=A-X$ and let imT_A denote its image. T_A is clearly injective, so $|imT_A|=|G(n,\mathbb{F})|$. Moreover, our estimate above shows that

$$\begin{aligned}|imT_A\cap G(n,\mathbb{F})| &= |imT_A|+|G(n,\mathbb{F})|-|imT_A\cup G(n,\mathbb{F})|\\ &\geqslant 2|G(n,\mathbb{F})|-|M(n,\mathbb{F})|\\ &>1.\end{aligned}$$

So, there are at least two elements in $imT_A\cap G(n,\mathbb{F})$, say Y_1, Y_2 with

$$T_A(X_1)=Y_1 \quad\text{and}\quad T_A(X_2)=Y_2.$$

Both of these provide decompositions of $A=X_1+Y_1=X_2+Y_2$ and one provides a splitting into distinct elements of $G(n,\mathbb{F})$ unless $\mathbb{F}$ has characteristic 2. For if $X_1\neq Y_1$ we are done. Otherwise

$A = 2X_1 = 2Y_1$ and if also $X_2 = Y_2$ we deduce that $A = 2Y_2$. But $A = 2Y_1 = 2Y_2$ yields the contradiction $Y_1 = Y_2$ unless $\mathbb{F}$ has characteristic 2, in which case lack of distinctness forces A to be the zero matrix.

(It is worth noting that something of the flavour of this proof can be recaptured over the infinite fields $\mathbb{R}$ (and $\mathbb{C}$) by the following topological approach. $G(n,\mathbb{R})$ is an open, dense subset of $M(n,\mathbb{R})$ and imT_A, being essentially just a translate of $G(n,\mathbb{R})$, is also open and dense. Thus $G(n,\mathbb{R}) \cap imT_A$ shares these properties and supplies an open, dense set of candidates with which to split A in the required manner.)

We are however still left with the stubborn case $q = 2$. The necessity to adopt a fresh approach here will in fact provide an alternative proof for the previous cases as well. Our motivation comes from facing the question: do we need to split every matrix in $M(n,\mathbb{F})$ or can we manage with just splitting some suitably representative (or canonical) matrices? At this point recall (for example, from [**1**, p. 167]) that there are matrices P, Q in $G(n,\mathbb{F})$ such that $PAQ = \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix}$. Thus it is enough to effect a splitting of the matrices $\begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix}$ for $k = 1, 2, \ldots, n$. (The zero matrix has already been dealt with.) If $q > 2$ (or if $\mathbb{F}$ is infinite) we can choose any a in $\mathbb{F}$ with $a \neq 0, 1$ to give:

$$\begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix} = \left(\begin{array}{c|c} (1-a)I_k & 0 \\ \hline 0 & -aI_{n-k} \end{array}\right) + aI_n$$

with both matrices easily seen to be invertible and distinct. If $q = 2$ we consider the cases of even k and odd k separately. If $k = 2r$ is even then:

$$\begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix} = \left(\begin{array}{cc|c} I_r & I_r & 0 \\ I_r & 0 & \\ \hline 0 & & I_{n-k} \end{array}\right) + \left(\begin{array}{cc|c} 0 & I_r & 0 \\ I_r & I_r & \\ \hline 0 & & I_{n-k} \end{array}\right)$$

is a suitable splitting. If k is odd, it is enough to provide a splitting of I_3 since we can then abut the splitting of I_{k-3} given above. But

$$I_3 = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

clinches the matter. Finally, we consider uniqueness of the splitting. From our previous discussion, for there to be any chance of uniqueness q must be 2 and A must be nonzero. Uniqueness then forces the number of nonzero elements of $M(n,2)$ to be the same as the number of unordered pairs of distinct elements of $G(n,2)$. That is, $|M(n,2)| - 1 = \frac{1}{2}|G(n,2)|(|G(n,2)| - 1)$. The last equation can be quickly shown to hold only when $n = 2$.

Reference

[1] S. Lipschutz, Linear Algebra, McGraw-Hill, 1974.