(VI.B) Normal Form of a (Polynomial) Matrix

We define an algorithm associating to \( M \in M_n(F[\lambda]) \) a matrix in normal form,

\[ M \rightarrow nf(M), \]

using certain row and column operations. This is similar in spirit to the algorithm

\[ A \rightarrow \text{rref}(A) \]

we have used throughout this course for \( A \in M_n(F) \). Why introduce column operations? Because we lose a row operation when we are dealing with polynomials: we can only divide a row by a scalar (not a polynomial), and this makes taking rref of matrices with polynomial entries a lost cause. (There is no rref(\( \lambda I - A \)).) It also means that multiplying a row by \( x \) is not invertible (of course, you can undo it on that row, but there isn’t an elementary matrix for that “undo”).

Here are the operations we shall permit:

DEFINITION 1. The elementary row and column operations on \( M \in M_n(F[\lambda]) \) are the same as for \( M_n(F) \), except you can’t multiply a row or column by a polynomial. You may
(i) multiply a row (column) by a scalar (\( = \) an element of \( F \))
(ii) add a polynomial multiple of a row (column) to a different row (column)
(iii) swap two rows (columns).

The elementary matrices representing these operations turn out to be invertible (meaning: they have an inverse in \( M_n(F[\lambda]) \)), and moreover all invertible matrices \( \in M_n(F[\lambda]) \) are products of them
(just as in the $M_n(F)$ case).\(^1\) We have the following somewhat obvious

**Proposition 2.** The determinant of an elementary matrix $E \in M_n(F[\lambda])$ is a scalar.

If you believe they’re invertible this is automatic, since $\operatorname{det} E \cdot \operatorname{det}(E^{-1}) = 1$; if $\operatorname{det} E$ is a polynomial then this is impossible – there is no matrix $E^{-1}$ with polynomial entries and 1 over a polynomial for determinant.

**Definition 3.** If one passes from $M$ to $N$ using elementary row and column operations then $M$ and $N$ are simply called equivalent.

**The Algorithm.** Now let $M$ be any nonzero matrix with entries in $F[\lambda]$. Some notation:
- We say $g(\lambda) \mid M$ if it divides every entry of $M$
- $\ell(M) :=$ lowest degree of any nonzero (polynomial) entry of $M$ (if $M$ contains a nonzero scalar, say 3, then of course $\ell(M) = 0$)
- The “first” entry of $M$ with a certain property will just mean the first you come upon if you read $M$ like a page of a book.

Define an operation $(\ast)$ on $M$ as follows: say $m = m_{ij}$ is the “first” entry of $M$ with $\deg m = \ell(M)$ (it’s an entry of least degree); perform row/column swaps to bring it to the $(1,1)$ position. Using the division algorithm, write all other entries in the first column as $q_i m + r_i$, where $\deg r_i < \deg m$; subtract $q_i \times (1^{\text{st}} \text{ row})$ from the $i^{\text{th}}$ row (for $i = 2, \ldots, n$), to reduce these entries (in the $1^{\text{st}}$ column) to $r_i$. Do the same for the first row. This concludes the operation $(\ast)$.

The matrix now looks like
\[
\begin{pmatrix}
m & 0 & \cdots & 0 \\
0 & r_2 & \cdots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & \cdots & 0 & r_n
\end{pmatrix},
\]

\(^1\)The book by Hoffman and Kunze has formal proofs of these facts in §7.4.
all \( r 's \) of lower degree than \( m \). If they are not all zero then we have reduced \( \ell(M) \).

If we apply the algorithm (*) repeatedly

\[
\begin{align*}
M &= M_0 \rightarrow M_1 \rightarrow M_2 \rightarrow \cdots \\
M_0 &= (g_1, S_1)
\end{align*}
\]

we reach a matrix of the form

\[
\begin{pmatrix}
g_1 & 0 & \leftrightarrow & 0 \\
0 & \uparrow & S_1 \\
0 & \downarrow
\end{pmatrix} =: (g_1, S_1)
\]

in a finite number of steps, because we cannot continue to reduce

\[
\ell(M_0) > \ell(M_1) > \ell(M_2) > \ldots
\]

for very long. At the end of this process, divide \( g_1 \) by the coefficient of its highest power of \( \lambda \) to make it monic. (We still denote the result by \( g_1 \).) Call the whole sequence we have performed so far (**) .

If \( \deg g_1 > \ell(S_1) \) (i.e. \( g_1 \) is not of minimal degree in this matrix) then applying (**) again to \((g_1, S_1)\) will produce \((g_2, S_2)\) with \( \deg g_2 \leq \ell(S_1) < \deg g_1 \). (The reason: \( \exists \) entry \( s \in S_1 \) with lower degree than \( g_1 \) \( \implies \) (**) will begin by swapping this element [or another, of degree \( \leq \deg s \)] with \( g_1 \). From then on each application of (*) within (**) cannot increase the degree of the upper left-hand entry.) Since

\[
\deg g_1 > \deg g_2 > \ldots
\]

cannot continue forever, we eventually must reach \((g_k, S_k)\) having \( \deg g_k \leq \ell(S_k) \), i.e. such that \( g_k \) has the lowest degree in the matrix, excluding \( 0 's \).

However, \( g_k \) still may not divide all the entries of \( S_k \), even if it is of lower degree. If \( g_k \nmid S_k \) then let \( s \) be the first entry of \( S_k \) such that \( g_k \nmid s \), and use the division algorithm to write \( s = g_kq + r, \deg r < \deg g_k \). Add the column containing \( s \) to the first column, changing
the matrix to
\[
\begin{pmatrix}
g_k & 0 & \leftrightarrow & 0 \\
\vdots \\
g_kq + r & S_k \\
\vdots
\end{pmatrix},
\]
and subtract \( q \) times the first row from the row of \( s \), to obtain
\[
\begin{pmatrix}
g_k & 0 & \leftrightarrow & 0 \\
\vdots \\
r & S_k \\
\vdots
\end{pmatrix}.
\]
Since \( \deg r < \deg g_k \), applying (**) produces \((g_{k+1}, S_{k+1})\) such that \( \deg g_{k+1} \leq \deg r < \deg g_k \) (same argument as in the last paragraph). Continuing on as long as \( g_i \nmid S_i \) we have once again
\[
\deg g_k > \deg g_{k+1} > \deg g_{k+2} > \ldots
\]
and the process must terminate with \((g_k, S_k)\) such that \( g_k \mid S_k \).

We have produced from \( M \) , using a well-defined algorithm,
\[
\begin{pmatrix}
f_{(1)} & 0 & \leftrightarrow & 0 \\
0 \\
\uparrow & M^{(1)} \\
0
\end{pmatrix}
\]
with \( f_{(1)} \) a polynomial in \( \lambda \) dividing the entries of \( M^{(1)} \). Perform the whole sequence of steps again on \( M^{(1)} \) to get \( f_{(2)}, M^{(2)} \) (with \( f_{(2)} \mid M^{(2)} \) both still divisible by \( f_{(1)} \) (why?), and so on — until we have a diagonal matrix \( N \) with diagonal entries \( f_{(1)}, f_{(2)}, \ldots, f_{(n)} \). Thus \( N \) is a normal matrix (see §VI.A) and is equivalent to \( M \); we write \( N = nf(M) \).

**Uniqueness and Invariant factors.** If \( R \) (resp \( C \)) is the product of elementary matrices corresponding to the row (resp. column) operations performed in the computation, then the relationship is
\[
R \cdot M \cdot C = nf(M) = N.
\]
What if we used a different algorithm to put $M$ in normal form, say

$$'R \cdot M \cdot 'C = 'N?$$

Then according to the following proposition, $N$ and $'N$ are the same (getting *deja vu* yet?)

**Proposition 4.** There is exactly one matrix in normal form “equivalent” to a given matrix $M$.

So you don’t have to do things in the rigid order I specified above, to find $nf(M)$. The value of the rigid algorithm is that it has already proved the *existence* part of this proposition (“there is a $nf$ matrix equivalent to $M$”).

Recall the notation from §VI.A: $\delta_k(M) =$ monic $gcd$ of determinants of $k \times k$ submatrices of $M$. It is easy to show\(^2\) that these are invariant under row and column operations (ergo the terminology “invariant factors” for their ratios). So if $M$ and $N$ are equivalent then $\Delta_k(M) = \Delta_k(N)$. Moreover it is really easy to compute the invariant factors for

$$N = \begin{pmatrix} f_1(\lambda) & 0 \\ \vdots & \ddots \\ 0 & f_n(\lambda) \end{pmatrix}.$$  

Clearly, since all the $f_i$ are monic, $\delta_n(N) = \det(N) = f_1 \cdots f_n$. Next, the $gcd$ of the determinants of all $(n-1) \times (n-1)$ minors is simply $\delta_{n-1}(N) = f_1 \cdots f_{n-1}$. In general $\delta_k(N) = f_1 \cdots f_k$ and so $\Delta_k(N) = f_k$.

The most obvious conclusion you can draw from this, is the theorem from §VI.A: the diagonal entries of $nf(M)$ are the invariant factors $\Delta_k(M)$. Slightly more subtle is the uniqueness part of the proposition above: if both $N$ and $'N$ were equivalent to $M$ then they are equivalent to each other, and thus share the same invariant factors, and therefore have the same diagonal entries! Notice that the

\(^2\)See for instance the very readable proof in the book by Hoffman and Kunze, p. 260. But if you look farther in §7.4, beware of their normal form algorithm (the “proof” of Theorem 9 on pp. 257-8 contains a subtle flaw).
invariant factors are playing here very much the same role as (the standard basis of) the row space did, back in §§II.C-D, in our proof that there was exactly one rref matrix row-equivalent to a given matrix in \( M_n(F) \).

Now let \( M = \lambda I - A \). In general \( nf(M) \) is going to look like

\[
\begin{pmatrix}
1 & 0 \\
\vdots & \\
1 & h_1(\lambda) \\
0 & \ddots & \ddots \\
& & & & h_r(\lambda)
\end{pmatrix} = R \cdot (\lambda I - A) \cdot C
\]

where \( h_r(\lambda) = m_A(\lambda) \). Taking determinants of both sides, since \( \det R \) and \( \det C \) are scalars, say \( \det R \cdot \det C = k \in F \), we have

\[ h_1(\lambda) \cdot \ldots \cdot h_r(\lambda) = k \cdot \det(\lambda I - A) = k \cdot f_A(\lambda). \]

Since the degree of the right-hand side = \( n \),

\[ \sum \deg(h_i(\lambda)) = n. \]

This is one way you can check you’ve done everything right.

**Remark 5.** Since \( f_A \) and \( h_1, \ldots, h_r \) are all monic polynomials, \( k = 1 \). So we’ve proved directly that the characteristic polynomial of \( A \) is the product of the (diagonal) entries of \( nf(\lambda I - A) \), i.e. \( \prod h_i(\lambda) = f_A(\lambda) \) (Corollary VI.A.10).

**Example 6.** We compute \( nf(\lambda I - A) \) for

\[
A = \begin{pmatrix}
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1
\end{pmatrix},
\]

by applying the algorithm to

\[
\lambda I - A = \begin{pmatrix}
\lambda - 1 & -1 & -1 \\
-1 & \lambda - 1 & -1 \\
-1 & -1 & \lambda - 1
\end{pmatrix} \rightarrow
\]

Note that $m_A(\lambda) = \lambda^2 - 3\lambda$ is exactly the minimal polynomial we had found before, while $\lambda \cdot (\lambda^2 - 3\lambda) = \lambda^2(\lambda - 3) = f_A(\lambda)$.

Exercises

(1) For

$$A = \begin{pmatrix} 7 & 12 & -12 \\ -2 & -3 & 4 \\ 2 & 4 & -3 \end{pmatrix},$$

compute $nf(\lambda I - A)$ via the algorithm above, and use it to determine $m_A$ and $f_A$.

(2) For

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -2 & -2 & 0 & 1 \\ -2 & 0 & -1 & -2 \end{pmatrix},$$

determine the invariant factors of $\lambda I - A$ by putting the latter in normal form.

(3) Prove that $A \in M_n(\mathbb{C})$ is diagonalizable if and only if $m_A$ has no repeated roots, via the following steps:

(a) Show that similar matrices have the same minimal polynomial.

(b) Show that the minimal polynomial of a diagonal matrix is the product of the $(\lambda - \lambda_i)$ where $\{\lambda_i\}$ are the distinct eigenvalues. (With (a), this gives the “only if” part.)

Henceforth assume $m_A$ has no repeated roots.

(c) Show that if $A$ has eigenvalue $\lambda$, then $p(A)$ has eigenvalue $p(\lambda)$, for any polynomial $p$. 
(d) Using part (c), prove that $m_A$ is equal to the product $\prod (\lambda - \lambda_i)$ over distinct eigenvalues of $A$. [Hint: First show that this product divides $m_A$. Note that $A$ is not assumed diagonalizable!]

(e) Now the nullity of $\prod (A - \lambda_i I)$ is $\leq$ the sum of the nullities of $(A - \lambda_i I)$ (why?). The latter nullities are the geometric multiplicities of the eigenvalues. Using this, show that $A$ is diagonalizable.