Cayley-Hamilton and related matters

Recall that the linear transformation $T : V \to V$ that is the 0-vector of $\mathcal{L}(V)$ (let’s follow the book’s notation and write $T_0$ for it) is the one for which $T_0(v) = 0_V$ for all $v \in V$.

There is good reason to want to know about the polynomial equations that a given transformation $T : V \to V$ satisfies. Here’s what that means.

a) The vector space $\mathcal{L}(V)$ has the additional structure of a ring, a notion under discussion **right now** in Abstract Algebra I (401), as one can take the composition of linear transformations to be a “multiplication” that must obey certain rules (it’s just Theorem 2.10 on page 87). In particular, we can compose a transformation $T$ with itself, obtaining something rightfully called $T^2$.

b) We should know by now that if $V$ is finite-dimensional of dimension $n$ over the field $\mathbb{F}$, specifying an ordered basis $\beta$ of $V$ determines an isomorphism

$$[ \beta ] : \mathcal{L}(V) \to M_{n \times n}(\mathbb{F}); \quad \text{here, } ([ \beta ](T)) \text{ is what we called } [T]_\beta \text{ earlier.}$$

This isomorphism takes a product $TU \in \mathcal{L}(V)$ to $[T]_\beta[U]_\beta = [TU]_\beta \in M_{n \times n}(\mathbb{F})$.\footnote{One says that $[ ]_\beta$ is an isomorphism of rings.}

In other words, “everything” in linear algebra in finite-dimensional vector spaces can always be done in terms of matrices and column vectors (which must be quite a relief!); to do so, you must first choose an ordered basis of $V$. It is also part of the story to remember what happens when you change to another basis, but the formula for that is easy to remember.

c) Let $g \in P(\mathbb{F})$, and write it out as $g(t) = \sum_{k=0}^d c_k t^k$, with all $c_k \in \mathbb{F}$. Then for a matrix $A \in M_{n \times n}(\mathbb{F})$, $g(A) = \sum_{k=0}^d c_k A^k$, where we make the canonical convention that $A^0$ is another symbol for the identity matrix $I_n$. For instance, if $g(t) = t^3 + t - 2$ and $A$ is an $n \times n$ matrix, then $g(A) = A^3 + A - 2I_n$. Check that for all polynomials $g_1(t)$ and $g_2(t)$, the matrices $g_1(A)$ and $g_2(A)$ commute.

Fix $n \in \mathbb{N}$. Given an $n \times n$ matrix $A$, we are seeking [non-zero] polynomials $g(t)$ for which $g(A) = T_0$. There always exist such $g$. Being the patient creatures that we are, we first just gaze at the set of all such polynomials $g(t)$, i.e., let

$$S_A = \{ g \in P[t] \mid g(A) = T_0 \}.$$  

What does one see?

1. First $S_A$ is a subspace of the vector space $P(\mathbb{F})$. Indeed, it is the nullspace (kernel) of the linear transformation

$$e_A : P(\mathbb{F}) \to M_{n \times n}(\mathbb{F}) \quad \text{given by } \quad e_A(g) = g(A)$$

(evaluation at $A$). We seek a non-trivial element of $\mathfrak{N}(e_A)$. 

\begin{align*}
(*) \quad e_A : P(\mathbb{F}) & \to M_{n \times n}(\mathbb{F}) \\
& \text{given by } e_A(g) = g(A)
\end{align*}
2. Of course, if we wish, we can restrict ourselves to $g$ lying in the familiar subspace $P_d(\mathbb{F}) = \{g \in P(\mathbb{F}) \mid \deg(g) \leq d\}$ of $P(\mathbb{F})$, for any $d \geq 0$, and we get the restrictions

$$(e_A)_d : P_d(\mathbb{F}) \to M_{n \times n}(\mathbb{F}),$$

all given, of course, by the formula in (*).

3. Note that our previous experience gives us that $P_d(\mathbb{F}) \subset P_{d+1}(\mathbb{F})$, $P_d(\mathbb{F})$ has its standard basis \{1, t, t^2, ..., t^d\} so $\dim P_d(\mathbb{F}) = d + 1$, and $\mathfrak{r}((e_A)_d) = \mathfrak{r}(e_A) \cap P_d(\mathbb{F})$. And also, $\dim M_{n \times n}(\mathbb{F}) = n^2$.

Why is that useful? Because by rank-nullity, the linear transformation $(e_A)_{n^2}$ must have non-trivial nullspace, as the dimension of $\dim P_{n^2}(\mathbb{F})$ is greater than that of $M_{n \times n}(\mathbb{F})$. In other words, $A$ satisfies some non-trivial polynomial equation of degree $\leq n^2$, i.e., $S_A$ contains such a polynomial.

4. Surely we can do better than that! Let $g(t) = f_A(t) = \det(A - tI)$ be the characteristic polynomial of $A$. We know that the solutions for the scalar $t$ in $f_A(t) = 0$ are the eigenvalues of $A$, etc., etc. It would be tempting to just plug in the matrix $t = A$ into $f_A(t)$ to get

$$f_A(A) = \det(A - AI) = \det T_0 = 0. \text{ Violà!}$$

Unfortunately, the above is bullshit—try to carry it out as stated for the general $2 \times 2$ matrix; you have to install a $2 \times 2$ matrix as an entry of a $2 \times 2$ matrix, ....

However, it is nonetheless true that $f_A \in S_A$ for all matrices $A$. This is the famous Cayley-Hamilton Theorem. In the case when $A$ is a diagonalizable $n \times n$ matrix with distinct eigenvalues, it is not all that difficult to see that this is correct.

But some matrices “refuse” to diagonalize. The correct proof of Cayley-Hamilton is based on breaking $\mathbb{F}^n$ down into subspaces that are invariant under $A$. Details suppressed for now, but I’ll add that a systematic treatment of invariant subspaces is the key to identifying the “simplest form” of a general matrix.

5. For a diagonalizable $n \times n$ matrix with $n$ distinct eigenvalues, it happens to be true that $f_A$ is “the best that one can do”. What does that mean? Rather elementary ring theory tells us that every matrix $A$ has what’s called its minimal polynomial, the polynomial $g$ of lowest degree—we must normalize the situation by assuming that its leading coefficient is 1 $\in \mathbb{F}$—for which $g(A) = T_0$. For a diagonalizable $n \times n$ matrix with $n$ distinct eigenvalues, $(-1)^n f_A$ is the minimal polynomial. It is a sound idea to start wondering how and when $(-1)^n f_A$ can fail to be the minimal polynomial. An easy observation to start with is that for $A = cI_n$, $f_A(t) = (c - t)^n = (-1)^n(t - c)^n$, but the minimal polynomial is just $(t - c)$. How about our little demon, a non-diagonalizable matrix in $M_{2 \times 2}(\mathbb{C})$ (the field is OK, but our friend is not so nice)?