Part I: A Plane Tale.

1. Addition and Multiplication. What kind of algebraic games can be played with 2×2 matrices? First of all, they can be added together and scaled in the most obvious manner:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} + \begin{bmatrix} x & u \\ y & v \end{bmatrix} = \begin{bmatrix} a+x & b+u \\ c+y & d+v \end{bmatrix} \quad \text{and} \quad \alpha \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} \alpha a & \alpha b \\ \alpha c & \alpha d \end{bmatrix}.$$

For these operations, a matrix is just like any string of numbers. The rectangular organization into rows and columns does not come in. *Multiplication* of matrices is another story, however, the product of two matrices being defined by:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x & u \\ y & v \end{bmatrix} = \begin{bmatrix} ax+by & au+bv \\ cx+dy & cu+dv \end{bmatrix}.$$

Note that the entry in row i and column j of the product AB is obtained by amalgamating the i-th *row* of the first factor A with the j-th *column* of the second factor B. We could also say that the k-th column of AB is obtained by letting A act on the k-th column of B.

This operation obeys most of the usual laws of multiplication, such as $A(B+C) = AB + AC$ and $(A+B)C = AC + BC$ ("distributivity") as well as $A(BC) = (AB)C$ ("associativity"), but it does not allow you to interchange the order of factors: $AB \neq BA$ in general. Another fact to be wary of is that $AB = 0$ is possible even if $A \neq 0$ and $B \neq 0$.

A hard look at the definition of the product should convince you that the two distributive laws do in fact hold. Associativity is not quite as obvious. One way to see it is to imagine every matrix expressed in terms of the so-called matrix units, which are matrices having a single entry $= 1$ and the others $= 0$. Any matrix can written as a linear combination (i.e. a sum of scalar multiples) of these, e.g.

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = a \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + b \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} + c \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} + d \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

The matrix unit having the 1 in row i and column j (and 0 elsewhere) is denoted by E_{ij}. It is a good exercise in matrix multiplication to check that $E_{ij}E_{kl}$ equals E_{il} if $j = k$, and yields the zero-matrix otherwise. Multiplying once more, we see that both $E_{ij}(E_{kl}E_{mn})$ and $(E_{ij}E_{kl})E_{mn}$ give the same result (what is it?). To verify the general associative law $A(BC) = (AB)C$, we imagine all three matrices expressed as linear combinations of matrix units. Repeated use of the distributive law now reduces both sides of this equation to the special cases already checked. Playing with matrix units also reveals some of the quirks of matrix multiplication. For instance, $E_{11}E_{12} = E_{12}$ and $E_{12}E_{11} = 0$ show that factors cannot normally be interchanged and that a product of non-zero matrices can be zero.

When, in the sequel, we wish to refer to the entries of a 2×2 matrix A, we shall use one of the following two notations:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad \text{or} \quad A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

The second of these can be confusing to the eye and will be used rarely. However it is obviously more systematic than the first. As an exercise you should once in your life slog through the verification that $A(BC) = (AB)C$, with these matrices having the entries a_{ij}, b_{kl}, c_{mn}.

2. The Inverse and Singularity. The matrix product is especially simple if one of the factors is a *diagonal* matrix $D(\alpha,\beta)=\alpha E_{11}+\beta E_{22}$:

$$\begin{bmatrix}\alpha & 0\\ 0 & \beta\end{bmatrix}\begin{bmatrix}a & b\\ c & d\end{bmatrix}=\begin{bmatrix}\alpha a & \alpha b\\ \beta c & \beta d\end{bmatrix}\quad\text{or}\quad\begin{bmatrix}a & b\\ c & d\end{bmatrix}\begin{bmatrix}\alpha & 0\\ 0 & \beta\end{bmatrix}=\begin{bmatrix}\alpha a & \beta b\\ \alpha c & \beta d\end{bmatrix}.$$

Note: left multiplication by a diagonal scales the rows, while right multiplication by a diagonal scales the columns, of the other matrix. If $\alpha=\beta=1$, we have the *identity matrix* $I=D(1,1)$ which, as a multiplier, has no effect at all: $IA=AI=A$ for any A.

The obvious equation

$$\begin{bmatrix}a & b\\ c & d\end{bmatrix}\begin{bmatrix}d & -b\\ -c & a\end{bmatrix}=\begin{bmatrix}ad-bc & 0\\ 0 & ad-bc\end{bmatrix}\tag{†}$$

turns out to be a real gold mine of information. The second factor appearing in it is visibly concocted from the first factor A, and is known as its *adjoint*, denoted $A^\star$. The scalar $(ad-bc)$ on the right hand side is called the *determinant*, $\det A$. Our equation (†) can thus be abbreviated as $AA^\star=(\det A)I$. You should check that $A^\star A=(\det A)I$, as well. Here is what we can deduce:

(1) If $\det A\neq 0$, the matrix A has an *inverse*

$$A^{-1}=\frac{1}{\det A}A^\star,\tag{‡}$$

which is a matrix with the property that $AA^{-1}=A^{-1}A=I$.

(2) If $\det A=0$, then A is *singular*, which means that there is a $B\neq 0$ with $AB=0$.

(3) Clearly, a singular A *cannot* have an inverse (multiplying $AB=0$ by A^{-1} would force $B=0$). *Conclusion:* every A is *either* invertible *or* singular, and the determinant says which.

(4) Substituting the right hand side of $A^\star=(a+d)I-A$ in (†) yields the *Cayley-Hamilton relation*

$$A^2-(a+d)A+(ad-bc)I=0,$$

which will be put to work in lesson 5. The scalar $(a+d)$ is called the *trace* of A.

The following items not directly related to (†) should also be noted

(5) Inversion reverses the order of factors, i.e. $(AB)^{-1}=B^{-1}A^{-1}$, because $ABB^{-1}A^{-1}=AIA^{-1}=I$.

(6) Looking back to the product AB, as given in lesson 1, and computing its determinant, we obtain $(ax+by)(cu+dv)-(cx+dy)(au+bv)=adxv+bcyu-bcxv-adyu=(ad-bc)(xv-yu)$, i.e. the miraculous formula:

$$\det(AB)=(\det A)(\det B).$$

One of the main points in this list is item (3), the dichotomy between invertible matrices and singular ones. We conclude this lesson by putting yet another wrinkle on it. Obviously A is singular if and only if there is a column $X\neq 0$ such that $AX=0$ (if necessary we can use X to make up a square matrix B such that $AB=0$). As shown by the right hand term of the equation

$$\begin{bmatrix}a & b\\ c & d\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}ax+by\\ cx+dy\end{bmatrix}=x\begin{bmatrix}a\\ c\end{bmatrix}+y\begin{bmatrix}b\\ d\end{bmatrix},\tag{$*$}$$

this means that one of columns of A must be a scalar multiple of the other. Under those circumstances the two columns are said to be *dependent*. Thus we have three different ways of recognizing the invertiblity of a matrix A: non-singularity, non-zero determinant, non-dependent columns.

3. Matrices as Transformations. This lesson is about pictures and is best understood with the help of sketch-pad and pencil.

Recall that, by means of a cartesian coordinate system, any pair (x, y) of real numbers can be regarded as a *point* in the plane. A more dynamic interpretation of the same pair (x, y) is that of a *vector*, i.e. an imaginary shift "x units east and y units north" applicable to any point. This has the advantage of giving geometric meaning to addition (perform one shift after the other) and to scaling (lengthen or shorten the shift by a factor). A vector is often depicted by an arrow showing the shift applied to a suitable point (by default, the origin). We will not be fussy about interpreting $X = (x, y)$ as either a point or a vector depending on circumstances.

Given a 2×2 matrix A, we now want to describe how multiplication by A *transforms* the points of the plane. We think of A as changing every point X to the "image point" AX. How is AX related to X?

As a vector, X can be written as $X = xI^{(1)} + yI^{(2)}$, where $I^{(k)}$ stands for the k-th column of the identity matrix. AX is directly given by the definition

$$AX = xA^{(1)} + yA^{(2)}, \tag{1}$$

where $A^{(k)}$ denotes the k-th column of A. Thus AX should be composed from the columns of A in the same way that X is made from the columns of I: if you picture X as a diagonal arrow in a vector rectangle with sides $xI^{(1)}$ and $yI^{(2)}$, the vector AX appears as the diagonal in the parallelogram with sides $xA^{(1)}$ and $yA^{(2)}$.

In fact, if you imagine a rectangular coordinate grid overlaid on the plane (as on graph paper), you will see it transformed into a (usually) skew grid made from copies of the parallelogram given by $A^{(1)}$ and $A^{(2)}$. Every X is simply transformed into the AX having the same "coordinates" x and y with respect to the new grid.

Many useful transformations of the plane are unrelated to matrices (for instance, the one which gives rise to logarithmic graph paper). How, then, can we tell whether a transformation T of the plane comes from a matrix A? Answer: T must be *linear*, i.e. it must satisfy

$$T(V + W) = T(V) + T(W) \qquad \text{and} \qquad T(\alpha W) = \alpha T(W). \tag{2}$$

For $X = (x, y)$, this obviously implies that $T(X) = xT(I^{(1)}) + yT(I^{(2)})$ (see?), which is the same as multiplying X by the matrix A whose columns are $T(I^{(1)})$ and $T(I^{(2)})$. Geometrically the linearity relations (2) say that T leaves the origin fixed, and preserves parallelism and collinear proportion.

For illustration we describe the action of three special kinds of matrices:

$$D(\alpha, \beta) = \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix} \qquad G_{12}(\gamma) = \begin{bmatrix} 1 & \gamma \\ 0 & 1 \end{bmatrix} \qquad \text{and} \quad R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

$D(\alpha, \beta)$ affects the coordinates of a point separately, multiplying each by the appropriate factor α or β. The coordinate axes stay put (though points on them may move), the grid remains rectangular (though it may get stretched or squeezed). If $\alpha \neq \beta$, any vector not parallel to one of the axes is tilted toward the axis with the stronger factor.

$G_{12}(\gamma)$ produces a "shear", which simply moves every point (x, y) horizontally by the amount γy. Every coordinate square is deformed into a parallelogram with the same base and height. Of course $G_{12}(\gamma)$ has a sister $G_{21}(\gamma)$ with the γ in the lower left corner, which moves points vertically in a similar mannner.

$R(\theta)$ acts as a rotation of the plane through the angle θ. Since its columns are the images of $I^{(1)}$ and $I^{(2)}$, the definition of sine and cosine shows that these vectors are both rotated through θ. But then, so are $xI^{(1)}$ and $yI^{(2)}$, and therefore the whole rectangle spanned by these.

4. Eigenvectors and Eigenvalues. Like *sauerkraut* and *wanderlust*, the German word *eigen* has been adopted by the English language because it does not have an exact equivalent here; its meaning hangs somewhere between "own", "proper", and "characteristic". To study the most characteristic properties of a square matrix A, there is no more revealing exercise than the tracking down of its "own" vectors.

A 2-column $V \neq 0$ is an *eigenvector* of A if multiplication by A does not twist it out of alignment but only scales it by some numerical value λ:

$$AV = \lambda V \qquad \text{or} \qquad (A - \lambda I)V = 0. \tag{1}$$

Note that an eigenvector never comes alone. If $AV = \lambda V$ then also $AW = \lambda W$ for every scalar multiple $W = \alpha V$. Hence the whole vector-line determined by V (i.e. all vectors of the same direction) experiences the action of A as a simple scaling by the *eigenvalue* λ.

There are several ways of computing eigenvalues and eigenvectors; any well-stocked computer library will contain more than one program dedicated to this task. For 2×2 matrices they are easily obtained via determinants as follows.

Since V must be $\neq 0$ to qualify as an eigenvector, equation (1) says that λ is an eigenvalue of A if and only if the matrix $(A - \lambda I)$ is singular, i.e. $\det(A - \lambda I) = (a - \lambda)(d - \lambda) - bc = 0$. This shows that an eigenvalue does not have much choice: it must satisfy the *characteristic equation*

$$\det(A - \lambda I) = \lambda^2 - (\operatorname{tr} A)\lambda + \det A = 0, \tag{2}$$

where $\operatorname{tr} A = a + d$ and $\det A = ad - bc$ are the *trace* and the *determinant* of A. This simplifies to

$$(\lambda - r)^2 = q \qquad \text{with} \qquad r = \frac{a+d}{2},\ q = \frac{(a-d)^2 + 4bc}{4}. \tag{3}$$

There are at most two solutions, namely $\lambda_{1,2} = r \pm \sqrt{q}$. Once these are determined, it is easy to solve $(A - \lambda I)V = 0$ for the actual eigenvectors V. The extreme case $A - \lambda I = 0$ needs no further clarification and will be excluded from the following discussion.

Three cases can occur:

(a) There might be two independent eigenvectors V_1 and V_2. Then we can combine the two equations $AV_k = \lambda_k V_k$ $(k = 1, 2)$ into a single one by writing $AM = MD$, where D is the diagonal matrix $D(\lambda_1, \lambda_2)$ and $M = \left[V_1, V_2\right]$ is a new matrix having the eigenvectors as columns (check this!). Since the independence of the latter makes M automatically invertible (cf. lesson 2), we can rewrite this as

$$D = M^{-1}AM \quad \text{or} \quad A = MDM^{-1}. \tag{4}$$

In this case A is said to be *diagonalizable*. It occurs when $q > 0$.

(b) The eigenvectors of A form a single vector-line. Of course, the effect of A then must be the same on every eigenvector. Therefore $\lambda_1 = \lambda_2$, and $q = 0$.

(c) There are no (real) * eigenvalues and eigenvectors, i.e. $q < 0$.

---

* If we are allowed to use *complex* numbers, even $q < 0$ leads to eigenvalues $\lambda_1 \neq \lambda_2$ and hence to a diagonalization $A = MDM^{-1}$ with *complex* matrices M and D. This works out because the theory developed here does not care about the *nature* of the numbers being processed but only about the feasibility of the required algebraic operations. The geometric imagery, however, would have to be taken much less literally in the complex environment.

5. Similarity and Standard Forms. The matrices A and B are said to be *similar* if there is an invertible matrix M such that

$$B = M^{-1}AM \quad \text{or} \quad A = MBM^{-1}. \tag{1}$$

It is easy to see that, if two matrices are similar to a third, they are similar to each other.

What about eigengadgets? If A and B are as above, we obviously have $(B - \lambda I) = M^{-1}(A - \lambda I)M$, and also $\det(B - \lambda I) = (\det M^{-1})\det(A - \lambda I)(\det M) = \det(A - \lambda I)$, by the multiplicative property of determinants (lesson 2, item 6).

In particular, *similar matrices have the same characteristic equation, hence the same determinant, trace, and eigenvalues.* They will not sport the same eigenvectors, in general, but they will fall into the same one of the patterns (a), (b), or (c) outlined in lesson 4. In fact, X is an eigenvector for B if and only if MX is one for A (check!).

Given any A, we shall now show a way of constructing an invertible matrix M such that the resulting $B = M^{-1}AM$ takes on one of the following three standard forms:

$$\text{(a)} \quad \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \qquad \text{(b)} \quad \begin{bmatrix} r & 1 \\ 0 & r \end{bmatrix} \qquad \text{(c)} \quad \begin{bmatrix} r & -s \\ s & r \end{bmatrix}. \tag{2}$$

In the diagonalizable case (a), the last lesson has already told us what to do: take the columns of M to be eigenvectors of A. In the other two cases we shall exploit the Cayley-Hamilton relation (lesson 2, item 4), which is just the characteristic equation with the matrix A itself substituted for λ, and which can therefore be rewritten as

$$(A - rI)^2 = qI, \tag{3}$$

with r and q as in lesson 4.

We set $M = \big[A_0Y, \alpha Y\big]$, where A_0 stands for $A - rI$. The column $Y \neq 0$ as well as the scalar $\alpha \neq 0$ will be fixed later. The trick is that left multiplication by A_0 scales and interchanges the columns of M:

$$A_0M = \big[A_0^2Y, A_0\alpha Y\big] = \big[qY, \alpha A_0Y\big], \tag{4$'$}$$

because of $A_0^2 = qI$. Let F (for "flip") be the matrix which has 0's on, and 1's off, the diagonal. Right multiplication by F swaps columns, i.e., $\big[V, W\big]F = \big[W, V\big]$ for any two columns V, W. Hence the computation (4$'$) can be continued as follows:

$$\ldots = \big[\alpha A_0Y, qY\big]F = MD(\alpha, q/\alpha)F, \tag{4$''$}$$

giving $A_0M = MD(\alpha, q/\alpha)F$. We now ensure the invertibility of M by taking Y to be some non-eigenvector of A_0 (see?), for instance, one of the columns of I. Multiplying by M^{-1}, we then have $M^{-1}A_0M = D(\alpha, q/\alpha)F$ and therefore $M^{-1}AM = rI + D(\alpha, q/\alpha)F$, that is:

$$M^{-1}AM = \begin{bmatrix} r & \alpha \\ q/\alpha & r \end{bmatrix}. \tag{4}$$

Comparing this with the forms (2), we obviously get the results (b) and (c) as advertised, provided we pick the scalar α as follows:

(b) If $q = 0$, put $\alpha = 1$. (c) If $q = -s^2 < 0$, put $\alpha = -s$.

6. Powers and Orbits. The similarity relation $A = MBM^{-1}$ is easily extended to matrix powers, that is to say

$$A^n = MB^nM^{-1}, \tag{1}$$

for any integer n. This follows at once from the obvious equation $(MBM^{-1})(MCM^{-1}) = MBCM^{-1}$, by putting $C = B, B^2, \ldots$, and so on. In view of lesson 5, we therefore have a means of evaluating the powers A^n, as soon as we can do it for the standard forms shown there. Here is what we get for the n-th power in the three cases:

$$\text{(a)} \quad \begin{bmatrix} \lambda_1^n & 0 \\ 0 & \lambda_2^n \end{bmatrix} \qquad \text{(b)} \quad \begin{bmatrix} r^n & nr^{n-1} \\ 0 & r^n \end{bmatrix} \qquad \text{(c)} \quad \rho^n \begin{bmatrix} \cos n\theta & -\sin n\theta \\ \sin n\theta & \cos n\theta \end{bmatrix},$$

where, for the last case, we have put $[r, s] = \rho[\cos\theta, \sin\theta]$. Case (a) is obvious, but some words of explanation may be in order for the other two. In case (b), the standard matrix is $B = rI + E_{12}$, and therefore $B^n = r^nI + nr^{n-1}E_{12}$ by an easy calculation based on the fact that $E_{12}^2 = 0$ (cf. lesson 1). After the change to "polar coordinates" ρ and θ, the standard matrix in case (c) can be written as $B = \rho R(\theta)$. At this point we have to dip into geometry: $R(\theta)^n$ is the n-fold iteration of a rotation through θ (cf. lesson 3), hence equals $R(n\theta)$.

In many applications one is interested in watching a point or vector being subjected to the same "transition matrix" A again and again. More precisely, one starts with an initial point $X_0 \neq 0$, and then generates a sequence $X_1, X_2, \ldots X_n, \ldots$, called the "orbit" (under A) of X_0, by the recipe

$$X_n = AX_{n-1} \qquad \text{so that} \quad X_n = A^nX_0. \tag{2}$$

To study the evolution of the orbit X_n, let us write $X_0 = MY_0$, where M is such that $B = M^{-1}AM$ is one of the standard forms. If we unleash A^n on this, we get $A^nMY_0 = MB^nY_0$. In other words:

the A-orbit $X_n = A^nX_0$ is the M-image of the B-orbit $Y_n = B^nY_0$.

For the sake of clarity, it is best to imagine the points $X_n = (x_n, y_n)$ and $Y_n = (u_n, v_n)$ plotted in two different planes, the x, y-plane and the u, v-plane (concretely represented by two different sheets of paper). We first look at the standard orbits Y_n in the u, v-plane and later transform them via M.

To avoid petty complications, we shall assume that A (and hence B) is non-singular and has no negative eigenvalues; this is automatic if the matrix $A - I$ is small, so that successive points in an orbit are not too far apart. In cases (a) and (b), we get $(u_n, v_n) = (\lambda_1^n u_0, \lambda_2^n v_0)$ and $(u_n, v_n) = r^{n-1}(ru_0 + nv_0, rv_0)$, respectively. Assuming that $\lambda_1 > \lambda_2$ in the former, we see that the slopes tend to zero, i.e.,

$$\lim_{n\to\infty} \frac{v_n}{u_n} = 0 \tag{3}$$

in both cases, provided only that $u_0 \neq 0$ in case (a). Thus, every orbit Y_n which does not start on the v-axis eventually lines up with the u-axis. Transformed by M into the x, y-plane, this says that every orbit tends align itself with the λ_1-eigenline, unless, in case (a), it is stuck on the λ_2-eigenline, where $\lambda_1 > \lambda_2$.

Slogan: *orbits are attracted to the dominant eigenline.*

What about case (c), where there *is* no eigenline? Remember that $B = \rho R(\theta)$, so that $Y_n = \rho^n R(n\theta)Y_0$. If $\rho = 1$, this clearly moves around the origin on a circle. If $\rho > 1$, it also expands by the factor ρ each time it turns through the angle θ, and we are looking at an outward spiral. Similarly $\rho < 1$ gives a contracting spiral. Finally, M transforms the circles into ellipses in the x, y-plane and deforms the spirals accordingly. Therefore, if $\det A = \rho^2$ equals 1, the orbit X_n moves on some kind of ellipse. If $\det A > 1$, it spirals outward; if $\det A < 1$, inward.

7. Linear Dynamical Systems. Equation (2) of the preceding lesson is often called a (linear) *difference equation* because it can be written as

$$X_n - X_{n-1} = PX_{n-1}, \tag{1}$$

where $P = A - I$ is usually a smallish matrix, reflecting a relatively modest transition $X_{n-1} \mapsto X_n$. The totality of orbits X_n spawned by such an equation is known as a (discrete) "linear dynamical sytem". To explore these patterns from a geometric point of view, we shall again assume that P is small enough to make A have only positive eigenvalues (if any). To avoid stationary points, we also assume that 1 is not an eigenvalue. Finally, we exclude the case $A = \alpha I$ as uninteresting.

(a) If A is diagonalizable, let the numbering be chosen so as to make $\lambda_1 > \lambda_2$. There are two corresponding eigenlines Λ_1 and Λ_2, formed by the scalar multiples of $M^{(1)}$ and $M^{(2)}$, respectively. Any orbit starting on one of these, stays on it. Other orbits eventually align themselves with Λ_1, as we have seen in lesson 6. Going backward on an orbit means using A^{-1} instead of A as the transition matrix. Since this reverses the inequality between the eigenvalues, Λ_2 is the dominant eigenline for A^{-1}. Therefore, orbits tend to *come from* alignment with Λ_2 and *go toward* alignment with Λ_1.

This state of affairs still allows two fundamentally different patterns, depending on whether or not the eigenvalues lie on different sides of the value 1.

(a1) If they do, any orbit will move inward roughly along Λ_2 and outward roughly along Λ_1. It is as though it were first attracted by the origin, and then (as it got closer) repelled by it. This pattern is called a *saddle*.

(a2) If they do not (say, both are < 1), all orbits will actually converge to the origin obeying the rules of alignment outlined above. This is called a *node*. (If both are > 1, the picture is the same, but the flow is outward.)

(a1) (a2)

(b) If there is only one eigenline Λ_1, the analysis given under (a) still applies, only the orbits now come from the same direction that they go toward, thus making a kind of 180° turn. This is also called a *node*, and the flow can go either in or out, depending on the location of the eigenvalue: $r < 1$ or $r > 1$.

(c) The case of no eigenline was described in lesson 6. The pattern is a system of *spirals*.

A common source of difference equations is their continuous counterpart, namely (linear) *differential equations*:

$$dX = HXdt, \tag{2}$$

where H is a matrix, and dt stands for an infinitesimal increment of time, while dX denotes the corresponding change in X. As in the scalar case, their exact solution uses the exponential function (cf. appendix), but approximate solutions can be obtained by giving dt an actual small value δ and reading equation (2) as saying approximately that $X_n - X_{n-1} = PX_{n-1}$, where $P = \delta H$.

8. The Dot Product. The basic building block of matrix multiplication is the product

$$\begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} x \\ \\ y \end{bmatrix} = ax + by \tag{1}$$

of a row and a column. When dealing with vectors only, we usually do not wish to commit ourselves to a definite row or column lay-out of the coordinates, so we define the *dot-product* $V \bullet X$ of two vectors $V = (a, b)$ and $X = (x, y)$ by the same formula (1). We shall presently see that this algebraic operation has a surprising geometric interpretation.

To start with, we define the *norm* $|V|$ of a vector V by the simple but crucial equation

$$|V|^2 = V \bullet V, \tag{2}$$

and observe that the general dot-product can be computed from norms by the obvious formula

$$2V \bullet X = |V|^2 + |X|^2 - |V - X|^2. \tag{3}$$

We say that V and X are *orthogonal* to each other if $V \bullet X = 0$, or equivalently $|V - X|^2 = |V|^2 + |X|^2$. In particular,

$$V \bullet X = 0 \Longrightarrow |V - X| \geq |V|. \tag{4}$$

Obviously we can always find a scalar t such that $(X - tV) \bullet V = 0$, and clearly the vector tV is unique. It is called the *projection* of X onto V and given by

$$\text{proj}_V(X) = tV = \frac{X \bullet V}{V \bullet V} V, \tag{5}$$

if $V \neq 0$. For any scalar α, we can write $(X - \alpha V) = (X - tV) - (\alpha - t)V$ as a difference of orthogonal vectors, and therefore get $|X - \alpha V| \geq |X - tV|$, by the inequality (4). This means that, among all vectors of the form αV, the projection tV is the closest to X, in the sense that it minimizes $|X - \alpha V|$.

Now for the geometry. If our coordinate system is rectangular, a comparison of (1) and (2) shows that *norm* = *length* by the Theorem of Pythagoras. Moreover, if we represent the vectors V and X by arrows $\overrightarrow{OP}$ and $\overrightarrow{OQ}$, the vector $V - X = \overrightarrow{PQ}$ will lie along the third side of the triangle so formed (see?), and Pythagoras says that $|V - X|^2 = |V|^2 + |X|^2$ if and only if the angle POQ is a right angle. Hence formula (3) implies that *orthogonality* = *perpendicularity*. In the same setting, $\text{proj}_V(X) = tV$ would show up as an arrow $\overrightarrow{OP'}$ with P' chosen so as to make $OP'Q$ a right angle. Now, if ϕ denotes the *angle* between V and X, we have $\cos\phi = t|V|/|X|$, which turns into

$$\cos\phi = \frac{X \bullet V}{|V|\,|X|}, \tag{6}$$

by the formula (5) for tV. This shows that the dot-product governs angles as well as lengths.

Remark: Equations (2) through (6) as well as the arguments connecting them are valid *verbatim* for three or more dimensions, since they hinge only on Pythagoras and (2). This will be used to give a little more zest to the exercises for this lesson.

In conclusion, we use the dot-product to give a geometric interpretation of $\det A$, in terms of the parallelogram $\mathcal{P}(V, W)$ spanned by the columns V and W of A. First you should note that $ad - cb = (a, c) \bullet (d, -b)$ means that $\det A = V \bullet W^{\perp}$, where $W^{\perp} = R(-90°)W$. Therefore, if ϕ and ψ denote the angles V makes with W and $W^{\perp}$, respectively, we have $\cos\psi = \pm\sin\phi$. Hence the *area* of $\mathcal{P}(V, W)$ is

$$|V|\,|W|\sin\phi = \pm V \bullet W^{\perp} = \pm\det A. \tag{7}$$

9. Orthogonal and Symmetric Matrices. The *transpose* M^T of a matrix M is formed by writing the columns of M as rows (and vice versa). For our little 2×2-matrix A it simply interchanges the off-diagonal entries.

How does this operation interact with products? Observe: the (ij)-th entry of $(MN)^T$ is the dot-product of the j-th row of M = j-th column of M^T with the i-th column of N = i-th row of N^T; whence

$$(MN)^T = N^T M^T. \tag{1}$$

In any discussion of the metric properties of a matrix A (i.e. those relating to distances and angles), the transpose looms large, because of the fundamental identity

$$Y \bullet AX = A^T Y \bullet X, \tag{2}$$

valid for any pair X, Y of vectors. If we imagine these written as columns, their transposes would be rows, and $Y \bullet X$ is the matrix product of Y^T with X. Then (2) just expresses the associativity relation $(Y^T A)X = Y^T(AX)$.

The square matrix A is said to be *orthogonal* if $A^T = A^{-1}$, i.e. if $A^T A = I$. The latter equation obviously says that the columns of A are of length 1 and mutually perpendicular (see?). You should verify that products of orthogonal matrices are again orthogonal. You should also note that $A^T A = I$ implies $Y \bullet X = A^T AY \bullet X = AY \bullet AX$, and find a geometric meaning for this.

A 2×2 orthogonal matrix A does not have much choice as to its appearance. If its first column is some vector V of unit length, orthogonality demands that its second column be $\pm V^\perp$. Accordingly we have two types:

$$\text{(i)} \quad A = \begin{bmatrix} x & -y \\ y & x \end{bmatrix} \qquad \text{(ii)} \quad A = \begin{bmatrix} x & y \\ y & -x \end{bmatrix}, \tag{3}$$

both with $x^2 + y^2 = 1$, which often makes it convenient to set $x = \cos\theta$, $y = \sin\theta$. Obviously type (i) is the good old rotation $R(\theta)$. It has determinant 1 and no eigenvectors. Type (ii) has eigenvalues $+1, -1$ (see?), and since it is symmetric (cf. below) its eigenvectors are perpendicular to each other. It is called a *reflection*.

The matrix A is said to be *symmetric* if $A^T = A$. In our 2×2-case this simply says $b = c$. Let us again exclude the trivial case $A = \alpha I$. Then $(a-d)^2 + 4c^2 > 0$, and we know from lesson 4 that A will have (real) eigenvalues $\lambda_1 \neq \lambda_2$. If V_1 and V_2 are corresponding eigenvectors, we have

$$\lambda_1 V_1 \bullet V_2 = AV_1 \bullet V_2 = V_1 \bullet AV_2 = \lambda_2 V_1 \bullet V_2 \quad \Longrightarrow \quad V_1 \bullet V_2 = 0, \tag{4}$$

i.e., the eigenvectors of A are perpendicular to each other. If we choose them to be of unit length (and who can stop us?), the matrix $M = \begin{bmatrix} V_1, V_2 \end{bmatrix}$ will even be orthogonal, and $A = MDM^{-1} = MDM^T$ with D diagonal. Conversely, every matrix of the form MDM^T is obviously symmetric (see?).

Slogan: *symmetric matrices are orthogonally diagonalizable.*

Amazingly it turns out that *any* square matrix can be written as a product $A = SR$, of a symmetric S and a rotation R. To compute S in the 2×2-case, let the first column of R have unknown entries x, y and force

$$AR^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x & y \\ -y & x \end{bmatrix} = \begin{bmatrix} * & ay + bx \\ cx - dy & * \end{bmatrix} \tag{5}$$

to be symmetric by picking (x, y) in the intersection of the line $(c-b)x = (a+d)y$ with the circle $x^2 + y^2 = 1$.

10. Elimination. Before leaving 2×2-matrices, we shall replace our simple treatment of inverses and determinants by a more subtle one, which depends less on lucky formulas for $\det A$ and A^{-1} and will therefore generalize to 3×3 matrices and beyond. Starting from scratch, we retain from lesson 2 only the *definitions* of concepts, and ignore all the interconnections already discovered.

We begin by introducing the so-called *elementary* matrices, namely:

$$G_{12}(s) = \begin{bmatrix} 1 & s \\ 0 & 1 \end{bmatrix} \quad G_{21}(s) = \begin{bmatrix} 1 & 0 \\ s & 1 \end{bmatrix} \quad D_1(s) = \begin{bmatrix} s & 0 \\ 0 & 1 \end{bmatrix} \quad D_2(s) = \begin{bmatrix} 1 & 0 \\ 0 & s \end{bmatrix} \quad F = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

where $s \neq 0$ is variable. Actually $F = D_1(-1)G_{21}(1)G_{12}(-1)G_{21}(1)$ (check this!) is only listed for convenience as a separate entity.

These matrices are obviously invertible, their inverses being $G_{12}(-s)$, $G_{21}(-s)$, $D_1(s^{-1})$, $D_2(s^{-1})$, F, respectively. Note that we need not *compute* these inverses, we simply verify that they work; i.e. that $AA^{-1} = A^{-1}A = I$ in each case. Note further that any product of elementary matrices is automatically invertible and therefore cannot be singular (why?). Presently we shall see that the converse also holds:

Every non-singular matrix is a product of elementary ones.

In fact, we shall describe an explicit computational recipe, called *elimination*, which either reveals the singularity of a matrix A, or produces an equation $E_4E_3E_2E_1A = I$ with elementary matrices E_k. To follow this description it is essential to understand that left multiplication of A by an elementary matrix E has a very straightforward effect on the *rows* of A: if $E = G_{ij}(s)$, the effect is to add s times row j to row i; if $E = D_k(s)$, it is to multiply row k by s. Finally, F switches the rows. You should check these carefully before reading on.

Unless interrupted by an unmistakable symptom of singularity, we can always reduce A as follows:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \to \begin{bmatrix} a' & b' \\ 0 & d' \end{bmatrix} \to \begin{bmatrix} 1 & b'' \\ 0 & d'' \end{bmatrix} \to \begin{bmatrix} 1 & b'' \\ 0 & 1 \end{bmatrix} \to \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{1}$$

where each arrow stands for an elementary multiplier E_k. Here is how: If $a = c = 0$, then A is singular; quit. If $a = 0$ and $c \neq 0$, take $E_1 = F$; if $a \neq 0$, take $E_1 = G_{21}(-c/a)$. Next, take $E_2 = D_1(1/a')$. If $d'' = 0$, then A is singular (see?); quit. Otherwise let $E_3 = D_2(1/d'')$, and finally $E_4 = G_{12}(-b'')$. End.

You should experiment with some numerical examples and try your own variations on this scheme. The general idea is to whittle down the matrix by a string of elementary moves.

Determinants. For another approach to these we first check the simple formula $\det(EA) = \det E \cdot \det A$, which is valid for elementary E and any A. For $E = G_{12}(s)$ this follows from $(a+sc)d - (b+sd)c = ad - bc$, for $E = G_{21}(s)$ it works similarly, and for the others it is even easier. Its repeated application yields

$$\det(E_m \cdots E_2E_1A) = \det E_m \cdots \det E_2 \cdot \det E_1 \cdot \det A, \tag{2}$$

for any elementary sequence E_k and any A, whence

$$\det(BA) = \det B \cdot \det A, \tag{3}$$

for any non-singular B (see?). It is a good exercise to prove once more, by this kind of reasoning, that $\det B = 0$ if and only if B is singular, and that (3) holds always. Though it seems more laborious, this approach to determinants is ultimately more powerful than the quick magic of lesson 1, which depends too much on the smallness of the matrices considered so far.